

Fisher Information in Flow Size Distribution Estimation

Paul Tune, *Member, IEEE*, and Darryl Veitch, *Fellow, IEEE*

Abstract—The flow size distribution is a useful metric for traffic modeling and management. Its estimation based on sampled data, however, is problematic. Previous work has shown that flow sampling (FS) offers enormous statistical benefits over packet sampling but high resource requirements precludes its use in routers. We present Dual Sampling (DS), a two-parameter family, which, to a large extent, provide FS-like statistical performance by approaching FS continuously, with just packet-sampling-like computational cost. Our work utilizes a Fisher information based approach recently used to evaluate a number of sampling schemes, excluding FS, for TCP flows. We revise and extend the approach to make rigorous and fair comparisons between FS, DS and others. We show how DS significantly outperforms other packet based methods, including Sample and Hold, the closest packet sampling-based competitor to FS. We describe a packet sampling-based implementation of DS and analyze its key computational costs to show that router implementation is feasible. Our approach offers insights into numerous issues, including the notion of ‘flow quality’ for understanding the relative performance of methods, and how and when employing sequence numbers is beneficial. Our work is theoretical with some simulation support and case studies on Internet data.

Index Terms—Fisher information, flow size distribution, Internet measurement, router measurement, sampling.

I. INTRODUCTION

The distribution of *flow size*, that is the number of packets in a flow, is a useful metric for traffic modelling and management, and is important for security because of the role small flows play in attacks. As is now well known however, its estimation based on sampled data is problematic.

Currently, sampling decisions in routers are made on a per-packet basis, with only sampled packets being subsequently assembled into (sampled) flows. Duffield et al. [1] were the first to point out that simple *packet sampling* strategies such as ‘1 in N ’ periodic or i.i.d. (independent, identically distributed) packet sampling have severe limitations, in particular a strong flow length bias which allows the tail of the flow size distribution to be recovered, but dramatically obscures the details of small flows. They explored the use of TCP SYN packets to improve the resolution at the small flow end of the spectrum. Hohn et al. [2], [3] explored these difficulties further and pointed out that *flow sampling*, where the sampling decision is made directly on flows, resulting in all packets belonging to any sampled flows being collected, has enormous statistical advantages. However, flow sampling has not been pursued further nor found its way into routers, partly because it implies that lookups be performed on every packet, which is very resource intensive.

More recently, Ribeiro et al. [4] explored the use of TCP sequence numbers to improve estimation for TCP flows. The idea is that the presence of packets which are not physically sampled can be inferred by noting the increasing byte count given by the sequence number fields of sampled packets. By using the Fisher information as a metric of the effectiveness of sampling in retaining information about the original flow sizes, they showed that this helps greatly to ‘fill in the holes’ left by packet sampling. However, they did not address whether these techniques out-perform flow sampling (FS).

In this paper we revisit FS in the context of TCP flows. Our first contribution is to explain how the approach of [4] can be reformulated and extended to include FS. This provides a framework for our second contribution, proofs that FS outperforms existing methods by a large margin, though they themselves greatly improve upon simple packet sampling. Our results are rigorous, based on explicit calculation and comparison of the Fisher Information matrices of competing schemes. With the statistical reputation of FS thus reinforced, the challenge is to find methods which can somehow approach or approximate flow sampling in order to benefit from its information theoretic efficiency, but with lower resources requirements. We show how this can be done.

The computational problem for FS can be described as follows. To capture the variety present in traffic flows and to provide the raw material for a variety of current (and future) metrics, many flows must be sampled. This implies large flow tables which in turn implies the use of slower but cheaper DRAM rather than the faster but expensive SRAM [5]. However, DRAM is not fast enough to perform lookups for every packet, as required by a straightforward implementation of FS, for today’s high capacity links. The question then becomes, how can flow sampling be implemented using per-packet decisions, in other words using some form of *packet sampling*?

The main contribution of this paper is the introduction of *Dual Sampling* (DS), a hybrid approach combining the advantages of both packet and flow sampling. It is a two parameter sampling family which includes FS as a special case and allows FS to be approached continuously, enabling a tradeoff of sampling efficiency against computational cost. Computationally, it can be implemented via a modified form of two-speed or ‘dual’ packet sampling which circumvents the problem of slow DRAM. There is a cost in terms of wasted samples, but we show that this can be borne in high speed routers. Following [4], DS benefits from the use of TCP sequence numbers although it can also be used without them, and we provide insight into how and when they have an

The authors are with the Department of E&E Engineering, The University of Melbourne, Australia (Email: {lsptune@ee.,dveitch@}unimelb.edu.au).

impact. We show rigorously that DS outperforms the methods proposed in [4]. We also compare and contrast DS with the well known ‘Sample and Hold’ scheme [6]. We show that Sample and Hold performs quite well, though not as well as DS.

Finally, we introduce *SYN+SEQ+FIN*, another sampling method which enables flow sampling to be perfectly achieved (aside from errors in the mapping of byte to packet counts) at very low computational cost, well below that even of packet sampling. Its disadvantage is that it exploits the TCP FIN field, when not all TCP flows terminate correctly with a FIN packet.

With its explicit use of TCP protocol information in most cases, our work applies to TCP flows only. However, the ideas and results could apply to other kinds of flows provided that suitable substitutes could be found for connection startup (SYN), ‘progress’ (sequence numbers) and termination (FIN). TCP flows still constitute the overwhelming majority of traffic in the Internet.

The rest of the paper is organized as follows. Section II describes our sampling framework and derives the Fisher information matrix and its inverse explicitly. Section III defines the sampling methods and derives their main properties. Section IV compares the methods theoretically and derives further properties explaining their performance, and Section V explores in more detail how sequence numbers reduce estimation variance. Section VI introduces a simple model for computational cost and uses it to define and solve an optimization problem for sampling performance under constraints. Section VII applies the methods to real Internet data and shows that DS performs favorably with flow sampling in practice, and has better performance than Sample and Hold. A closed-form unbiased estimator was proposed for DS and Sample and Hold which achieves the Cramér-Rao lower bound asymptotically, eliminating the need for iterative optimization algorithms. We conclude and discuss future work in Section VIII.

This paper is an extended and enhanced version of the conference paper [7]. The main additions relate to the inclusion of the Sample and Hold method throughout the paper, several new theorems and counter-examples on method comparison, and the inclusion of a new major data set.

II. THE SAMPLING FRAMEWORK

In this section we establish a framework to define and analyze sampling techniques applied to an idealized view of TCP flows on a link. Nominally, we imagine that such flows are defined by the usual 5-tuple of origin and destination IP addresses, port numbers, and TCP protocol field together with a timeout. For the analysis we make a number of simplifying assumptions:

- (i) flows begins with a SYN packet and have no others,
- (ii) flows are not split (this can occur through timeouts or flow table clearing),
- (iii) all necessary protocol information (5-tuple, SYN/FIN bits and sequence numbers) can be observed, and
- (iv) per-flow sequence numbers count packets, not bytes.

Assumptions (iii) and (iv) will be discussed/relaxed when we deal with real data in Section VII. Note that we **do** respect

TCP’s per-flow random initialization of sequence numbers. Hence their absolute value holds no information on the number of packets in a flow, only differences of sequence numbers matter. This is crucial for the analysis.

A. The Flow Model

We consider a measurement interval containing N_f flows. Let m_i denote the *size* of flow i (the number of packets it contains). It satisfies $1 \leq m_i \leq W$, where $1 \leq W < \infty$ is the maximum flow size. The total number of packets is $n = \sum_{i=1}^{N_f} m_i$.

Let M_j be the number of flows of size j , $1 \leq j \leq W$, that is $M_j = \sum_{i:m_i=j} 1$, and $N_f = \sum_{j=1}^W M_j$. The flow size ‘distribution’ is the set $\theta = \{\theta_1, \theta_2, \dots, \theta_W\}$ of relative frequencies, that is

$$\theta_j = \frac{M_j}{N_f} \quad (1)$$

where $0 \leq \theta_j \leq 1$ and $\sum_{j=1}^W \theta_j = 1$. Note that $\{M_j\}$ and $\{\theta, N_f\}$ are equivalent and complete descriptions of the flows in the measurement interval. They are sets of deterministic *parameters*, not random variables, effectively a deterministic flow size model.

Most of the literature on traffic sampling follows the above viewpoint, where the data is deterministic, the only randomness being introduced through the sampling itself. An exception is the work of Hohn and Veitch ([2], [3]) where randomness arises both through the traffic model and the sampling process, which makes the analysis considerably more difficult, but less generic.

B. General Random Sampling

The effect on flow size of random sampling can be described as follows: from an original flow of size k , only j packets, $0 \leq j \leq k$, are *sampled* (retained), with probability

$$b_{jk} = \Pr(\text{sampled flow has } j \text{ pkts} \mid \text{original flow has } k \text{ pkts}).$$

The operation of the sampling scheme is entirely defined by the b_{jk} , which can be assembled into a $(W+1) \times W$ *sampling matrix* \mathbf{B} , whose $(j+1, k)$ -th element is b_{jk} . Note that $b_{jk} = 0$ for $j > k$. By definition \mathbf{B} is a (column) *stochastic matrix*, that is each element obeys $b_{jk} \geq 0$, and each column sums to unity.

The experimental outcome can be described by a set of random variables $\{M'_j \mid 0 \leq j \leq W\}$ where M'_j counts the number of sampled flows of size j . Thus,

$$M'_j = \sum_{i:m'_i=j} 1 = \sum_{i=1}^{N_f} \mathbb{1}(m'_i = j).$$

Equivalently, let $\theta' = \{\theta'_j \mid 0 \leq j \leq W\}$ (note that the index j includes 0) denote the empirical distribution of *sampled flow* sizes, where

$$\theta'_j = \frac{M'_j}{N_f}, \quad (2)$$

where $0 \leq \theta'_j \leq 1$ and $\sum_{j=0}^W \theta'_j = 1$. Note that $\theta'_0 \geq 0$ as some flows may not survive the thinning process. Out of the

original N_f flows, only $N'_f = N_f(1 - \theta'_0) = N_f - M'_0$ flows survive sampling.

Define a normalized set of fractions of the sampled flow sizes $\gamma = \{\gamma_j \mid 1 \leq j \leq W\}$ by

$$\gamma_j = \frac{\theta'_j}{\sum_{k=1}^W \theta'_k} = \frac{\theta'_j}{1 - \theta'_0}, \quad (3)$$

where $0 \leq \gamma_j \leq 1$ and $\sum_{j=1}^W \gamma_j = 1$. The set γ constitutes the naïve or directly measurable sampled flow size distribution and is equivalent to the distribution of flows conditional on at least a single packet from that flow being sampled. For $j \geq 1$, θ'_j is related to γ_j by $(1 - \theta'_0)\gamma_j = (N'_f/N_f)\gamma_j = \theta'_j$.

C. The Unconditional Formulation

The sampled flow above includes the case, $j = 0$, where the flow ‘evaporates’. It seems natural to conclude however that such cases cannot be observed. This logically leads to an analysis based on the observation of the γ_j defined above where $j \geq 1$, which is effectively *conditional*: sample flow distributions given that at least one packet is sampled. This is the approach adopted in [1], [4], [8] and in the literature generally. One of the key differences in our work is that we show that it is possible to observe the $j = 0$ case, leading to an *unconditional* formulation which enjoys many advantages.

To see how this is possible we return to general context of N_f flows, *each* one of which will be sampled in this general sense. As defined above N'_f is the number of flows of size at least 1 after sampling. The number of evaporated flows is just $N_f - N'_f$, but typically N_f is not known and is regarded as a ‘nuisance parameter’ which must be estimated. However, it can easily be measured by directly counting the total number of SYN packets, which is just the number of flows N_f . For methods which are already assuming an ability to access and perform specific actions based on whether a packet is a SYN or not, the additional assumption of being able to count all SYN packets is a natural one. It is also implementable, as a single additional counter which checks every packet and conditionally increments based on a small number of header bits is not difficult even at the highest speeds [5], as we discuss in more detail in Section VI. In summary, by knowing N_f , every flow gives rise to a sampled flow, each one of which is observable, either directly ($j \geq 1$), or indirectly ($j = 0$). In other words, we can in effect observe the θ'_j over the full range from $j = 0$ to $j = W$.

The chief advantage of the unconditional formulation is the very simple form of the likelihood function for the experimental outcome j for a single flow. This makes the manipulation of the Fisher information far more tractable, leading to new analytic results and insights. The other big advantage is that flow sampling can now be included. In the conditional world flow sampling is perfect – by definition, if a flow is sampled at all, all its packets will be and so there is nothing to do! The unconditional framework allows the missing part of the picture to be included, enabling meaningful comparison.

D. The Sampled Flow Distribution

Our analysis is based on the idea of selecting a ‘typical’ flow, and that flows are mutually independent (a reasonable

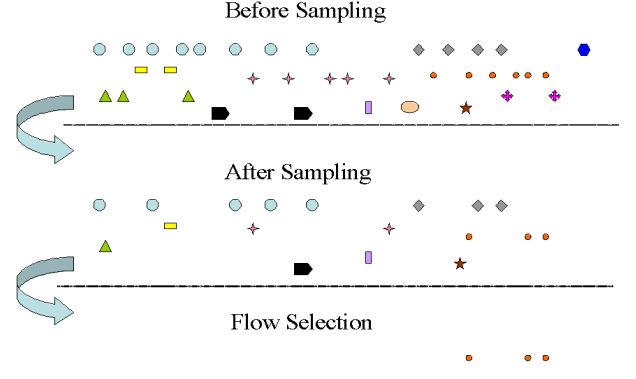


Fig. 1. The flow sampling and selection process. Here a flow is selected which had $k = 5$ packets originally and $j = 3$ after sampling.

assumption if N_f is very large). Since flows are in fact deterministic, this is only meaningful if we introduce a supplementary random variable U , a uniform over the N_f flows available, which performs the random flow selection. This variable, which acts ‘invisibly’ behind the scenes (and is rarely discussed), is not part of the random sampling scheme itself, but is essential as it allows the θ parameters to be treated as probabilities, even though they are not. An example is given in Figure 1 which shows $N_f = 12$ flows before and after sampling, followed by a random flow selection. In the interests of clarity a flow of size $j = 3$ after sampling was selected, but it could have been one of the evaporated flows ($j = 0$).

With this background established, the discrete distribution for a sampled flow originally of size k is very simple:

$$d_j = \sum_{k=1}^W b_{jk} \theta_k, \quad 0 \leq j \leq W. \quad (4)$$

This can be expressed in matrix notation as

$$\mathbf{d} = \mathbf{B}\boldsymbol{\theta} \quad (5)$$

where $\mathbf{d} = [d_0, d_1, d_2, \dots, d_W]^T$ is a $(W+1) \times 1$ column vector, and $\boldsymbol{\theta}$ a $W \times 1$ column vector. The probability d_j is related to the empirical fraction θ'_j for $j \geq 0$ by

$$\begin{aligned} \mathbb{E}[\theta'_j] &= \frac{\mathbb{E}[\sum_{i: m'_i=j} 1]}{N_f} = \frac{\sum_{i=1}^{N_f} \mathbb{E}[\mathbb{1}(m'_i=j)]}{N_f} \\ &= \frac{\sum_{i=1}^{N_f} \Pr(m'_i=j)}{N_f} = \frac{N_f d_j}{N_f} = d_j. \end{aligned}$$

The likelihood function for the parameters is simply

$$f(j; \boldsymbol{\theta}) = d_j, \quad 0 \leq j \leq W. \quad (6)$$

In the conditional framework commonly used $j = 0$ is missing, and normalization is then needed to ensure probabilities add to one. This implies a division of random variables, which greatly complicates the likelihood.

E. The Fisher Information of a Sampled Flow

The parameter vector $\boldsymbol{\theta}$ is the unknown we would like to estimate from sampled flows. Since here we are not concerned

with specific estimators of θ , but in the effectiveness of the underlying sampling scheme, a powerful approach (introduced in [4]) is to use the *Fisher information* [9, Section 11.10] to access its efficiency in collecting information about θ .

We first introduce notation that will be used throughout this paper. The expectation of a random variable X is denoted by $\mathbb{E}[X]$, and the variance by $\text{Var}(X)$. Matrices are written in bold-face upper case and vectors in bold-face lower case. The *transpose* of a matrix \mathbf{A} is denoted by \mathbf{A}^T . The operator also applies to vectors. The operator $\text{tr}(\mathbf{A})$ denotes the trace of the matrix \mathbf{A} . The matrix \mathbf{I}_n denotes the $n \times n$ identity matrix. The vector $\mathbf{1}_n = [1, 1, \dots, 1]^T$ denotes an $n \times 1$ vector of 1s. The vector $\mathbf{0}_n$ denotes the $n \times 1$ null vector, and the $m \times n$ null matrix is written as $\mathbf{0}_{m \times n}$. Given an $n \times 1$ vector \mathbf{x} , $\text{diag}(\mathbf{x})$ denotes an $n \times n$ matrix with diagonal entries x_1, x_2, \dots, x_n .

Definition 1: An $n \times n$ real matrix \mathbf{M} is *positive definite* iff for all vectors $\mathbf{z} \in \mathbb{R}^n \setminus \{\mathbf{0}_n\}$, $\mathbf{z}^T \mathbf{M} \mathbf{z} > 0$, and is *positive semidefinite* iff $\mathbf{z}^T \mathbf{M} \mathbf{z} \geq 0$.

We write $\mathbf{A} > 0$ or $0 < \mathbf{A}$ to indicate that \mathbf{A} is positive definite. For two matrices \mathbf{A} and \mathbf{B} , we write $\mathbf{A} > \mathbf{B}$ to mean $\mathbf{A} - \mathbf{B} > 0$ in the positive definite sense. Similarly, $\mathbf{A} \geq \mathbf{B}$ and $\mathbf{A} - \mathbf{B} \geq 0$ each mean that $\mathbf{A} - \mathbf{B}$ is positive semidefinite. The operator $|\cdot|$ returns the size of a vector or set. All other definitions will be defined when needed.

The Fisher information is useful because its inverse is the Cramér-Rao lower bound (CRLB), which lower-bounds the variance of any unbiased estimator of θ . In fact the Fisher information takes a different form depending on whether constraints are imposed on the θ or not [10]. Inequality constraints are particularly problematic, so we avoid them by assuming that each θ_k obeys $0 < \theta_k < 1$ (this ensures that the CRLB optimal solution cannot include boundary values, which would create bias and thereby invalidate the use of the unbiased CRLB). Assuming that flows exist for all sizes, i.e. that $\theta_k > 0$ for all k , is reasonable given the huge number of simultaneously active flows (up to a million) in high end routers. There is one more constraint, the equality constraint $\sum_{k=1}^W \theta_k = 1$, which must be included. As this complicates the Fisher information, we first deal with the unconstrained case.

F. The Unconstrained Fisher Information

The Fisher information is based on the likelihood and is defined by

$$\begin{aligned} \mathbf{J}(\theta) &= \mathbb{E}[(\nabla_{\theta} \log f(j; \theta))(\nabla_{\theta} \log f(j; \theta))^T] \\ &= \sum_{j=0}^W (\nabla_{\theta} \log f(j; \theta))(\nabla_{\theta} \log f(j; \theta))^T d_j. \end{aligned} \quad (7)$$

Here $\nabla_{\theta} \log f(j; \theta) = (1/d_j)[b_{j1}, \dots, b_{jW}]^T$ because of the simple form (6) of the likelihood. This leads to the simple explicit expression $(\mathbf{J}(\theta))_{ik} = \sum_{j=0}^W \frac{b_{ji} b_{jk}}{d_j}$, or equivalently

$$\mathbf{J}(\theta) = \mathbf{B}^T \mathbf{D}(\theta) \mathbf{B} \quad (8)$$

where $\mathbf{D}(\theta)$ is a diagonal matrix with $(\mathbf{D}(\theta))_{jj} = d_j^{-1}$.

We will need to find the inverse of \mathbf{J} , but since \mathbf{B} is not square, this cannot be done directly from (8) in terms of the inverse of \mathbf{B} . However if we re-express \mathbf{B} as

$$\mathbf{B} = \begin{bmatrix} \mathbf{b}_0^T \\ \tilde{\mathbf{B}} \end{bmatrix} \quad (9)$$

where $\mathbf{b}_0^T = [b_{01}, \dots, b_{0W}]$ is the top row of \mathbf{B} and

$$\tilde{\mathbf{B}} = \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1W} \\ 0 & b_{22} & \dots & b_{2W} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & b_{WW} \end{bmatrix},$$

then we can write

$$\mathbf{J}(\theta) = \frac{1}{d_0} \mathbf{b}_0 \mathbf{b}_0^T + \tilde{\mathbf{J}}(\theta) \quad (10)$$

where $\tilde{\mathbf{J}}(\theta) = \tilde{\mathbf{B}}^T \tilde{\mathbf{D}}(\theta) \tilde{\mathbf{B}}$, $\tilde{\mathbf{D}}(\theta) = \text{diag}(d_1^{-1}, \dots, d_W^{-1})$, and $d_0 = \mathbf{b}_0^T \theta$. Since $\tilde{\mathbf{B}}$ and $\tilde{\mathbf{D}}(\theta)$ are square, the inverse of $\tilde{\mathbf{J}}(\theta)$ is just $\tilde{\mathbf{J}}^{-1}(\theta) = \tilde{\mathbf{B}}^{-1} \tilde{\mathbf{D}}^{-1}(\theta) (\tilde{\mathbf{B}}^{-1})^T$, that is

$$(\tilde{\mathbf{J}}^{-1}(\theta))_{ik} = \sum_{j=1}^W b'_{ij} b'_{kj} d_j \quad (11)$$

where $b'_{jk} = (\tilde{\mathbf{B}}^{-1})_{jk}$. By Lemma 28(ii) in Appendix B, $\tilde{\mathbf{J}}(\theta)$ is positive definite. We can now give the inverse of $\tilde{\mathbf{J}}$.

Proposition 2: The inverse of $\mathbf{J}(\theta)$ is given by

$$\mathbf{J}^{-1}(\theta) = \tilde{\mathbf{J}}^{-1}(\theta) - \frac{1}{d_0 + \mathbf{b}_0^T \tilde{\mathbf{J}}^{-1}(\theta) \mathbf{b}_0} \tilde{\mathbf{J}}^{-1}(\theta) \mathbf{b}_0 \mathbf{b}_0^T \tilde{\mathbf{J}}^{-1}(\theta).$$

Proof: The matrix inversion lemma applies (see Lemma 30 in Appendix A) with $\mathbf{R} = \tilde{\mathbf{J}}$ and $\mathbf{T} = 1/d_0$ nonsingular. Since $d_0 + \mathbf{b}_0^T \tilde{\mathbf{J}}^{-1}(\theta) \mathbf{b}_0 > \mathbf{b}_0^T \tilde{\mathbf{J}}^{-1}(\theta) \mathbf{b}_0 > 0$ as $\tilde{\mathbf{J}}^{-1}$ is positive definite, the result immediately follows. ■ The diagonal elements of the matrix $\mathbf{J}^{-1}(\theta)$ will be important in later sections, and the explicit formula is given below:

$$(\mathbf{J}^{-1}(\theta))_{jj} = \sum_{k=1}^W b_{jk}^2 d_k - \frac{\left(\sum_{k=j}^W d_k b'_{jk} \sum_{\ell=1}^k b_{0\ell} b'_{\ell k} \right)^2}{d_0 + \sum_{k=1}^W d_k \left(\sum_{\ell=1}^k b_{0\ell} b'_{\ell k} \right)^2}. \quad (12)$$

Again, this explicit inverse, valid for any general sampling matrix \mathbf{B} , is made possible by the very simple form of the likelihood function in equation (6). We now specialize the above result for sampling matrices that satisfy particular conditions. Although some of these matrices exhibit a dependence on θ , we drop this dependence for notational simplicity when the context is clear.

Corollary 3: If for some constant q the sampling matrix \mathbf{B} satisfies $\mathbf{b}_0 = q \mathbf{1}_W$ then (setting $p = 1 - q$)

$$\mathbf{J}^{-1} = \tilde{\mathbf{J}}^{-1} - \frac{q}{p} \theta \theta^T.$$

Proof: The given condition implies that $d_0 = \mathbf{b}_0^T \theta = q \mathbf{1}_W^T \theta = q$, and $\mathbf{1}_W^T \tilde{\mathbf{B}}^{-1} = (1/p) \mathbf{1}_W^T$ since \mathbf{B} is column

stochastic. Next,

$$\begin{aligned} \mathbf{b}_0^T \tilde{\mathbf{J}}^{-1} \mathbf{b}_0 &= q^2 \mathbf{1}_W^T \tilde{\mathbf{J}}^{-1} \mathbf{1}_W = q^2 \mathbf{1}_W^T \tilde{\mathbf{B}}^{-1} \tilde{\mathbf{D}}^{-1} (\tilde{\mathbf{B}}^{-1})^T \mathbf{1}_W \\ &= (q^2/p^2) \mathbf{1}_W^T \tilde{\mathbf{D}}^{-1} \mathbf{1}_W \\ &= (q^2/p^2) \sum_{j=1}^W d_j = (q^2/p^2)(1 - d_0) = q^2/p. \end{aligned}$$

Let $\tilde{\mathbf{d}} = \tilde{\mathbf{B}}\boldsymbol{\theta} = [d_1, d_2, \dots, d_W]^T$. Then from Proposition 2,

$$\begin{aligned} \mathbf{J}^{-1} &= \tilde{\mathbf{J}}^{-1} - \frac{1}{d_0 + q^2/p} \tilde{\mathbf{J}}^{-1} \mathbf{b}_0 \mathbf{b}_0^T \tilde{\mathbf{J}}^{-1} \\ &= \tilde{\mathbf{J}}^{-1} - \frac{q^2/p^2}{q + q^2/p} \tilde{\mathbf{B}}^{-1} \tilde{\mathbf{D}}^{-1} \mathbf{1}_W \mathbf{1}_W^T \tilde{\mathbf{D}}^{-1} (\tilde{\mathbf{B}}^T)^{-1} \\ &= \tilde{\mathbf{J}}^{-1} - \frac{q^2/p^2}{q + q^2/p} \tilde{\mathbf{B}}^{-1} \tilde{\mathbf{d}} \tilde{\mathbf{d}}^T (\tilde{\mathbf{B}}^T)^{-1} \\ &= \tilde{\mathbf{J}}^{-1} - \frac{q}{p} \boldsymbol{\theta} \boldsymbol{\theta}^T. \end{aligned}$$

The matrix $\tilde{\mathbf{J}}$ corresponds to the information carried by the outcomes $1 \leq j \leq W$ only. We expect \mathbf{J} to carry more information through the knowledge of N_f which gives access to $j = 0$, and therefore \mathbf{J}^{-1} to have reduced uncertainty, corresponding (through the CRLB) to a reduced variance. The following result confirms this intuition (proof in Appendix A).

Theorem 4: An upper bound for $\mathbf{J}^{-1}(\boldsymbol{\theta})$ is

$$\mathbf{J}^{-1}(\boldsymbol{\theta}) \leq \tilde{\mathbf{J}}^{-1}(\boldsymbol{\theta}).$$

Equality holds if and only if $\mathbf{b}_0 = \mathbf{0}_{W \times 1}$.

The reduction in uncertainty is given by the second term in the expression for $\mathbf{J}^{-1}(\boldsymbol{\theta})$ in Proposition 2.

G. The Constrained Fisher Information and CRLB

Intuitively, constraints on the parameters should increase the Fisher information since they tell us something more about them, ‘for free’. In fact, [11] shows that this is only true for equality constraints. Since we are assuming that $0 < \theta_k < 1$, the only active constraint is $\sum_{k=1}^W \theta_k = 1$. Its *gradient* is

$$\mathbf{G}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}) \quad (13)$$

where $g(\boldsymbol{\theta}) = \sum_{j=1}^W \theta_j - 1$.

The inverse constrained Fisher information [11] is

$$\mathcal{I}^+ = \mathbf{J}^{-1} - \mathbf{J}^{-1} \mathbf{G} (\mathbf{G}^T \mathbf{J}^{-1} \mathbf{G})^{-1} \mathbf{G}^T \mathbf{J}^{-1} \quad (14)$$

where \mathcal{I}^+ denotes the Moore-Penrose pseudo-inverse [12, Chapter 20, pp. 493-514] of the constrained Fisher information matrix \mathcal{I} . The matrix \mathcal{I}^+ is rank $W - 1$ due to the single equality constraint and is thus singular (see [11, Remark 2]). This somewhat formidable expression can be simplified in our case, as we now show.

Lemma 5: $\mathbf{J} \mathbf{diag}(\theta_1, \dots, \theta_W) \mathbf{1}_W = \mathbf{1}_W$.

Proof: The row sum of row i of $\mathbf{J} \mathbf{diag}(\theta_1, \dots, \theta_W)$ is

$$\sum_{k=1}^W \sum_{j=0}^W \frac{b_{ji} b_{jk}}{d_j} \theta_k = \sum_{j=0}^W \frac{b_{ji}}{d_j} \sum_{k=1}^W b_{jk} \theta_k = \sum_{j=0}^W \frac{b_{ji}}{d_j} d_j = 1$$

since \mathbf{B} is column stochastic. ■

It is easy to see that $\mathbf{G} = \mathbf{1}_W$. Hence $\mathbf{J}^{-1} \mathbf{G} = \mathbf{J}^{-1} \mathbf{1}_W = \mathbf{diag}(\theta_1, \dots, \theta_W) \mathbf{1}_W = \boldsymbol{\theta}$ from Lemma 5. It is then straightforward to verify that $(\mathbf{G}^T \mathbf{J}^{-1} \mathbf{G})$ is simply the number 1, and further that (14) can be reduced to

$$\mathcal{I}^+ = \mathbf{J}^{-1} - \boldsymbol{\theta} \boldsymbol{\theta}^T. \quad (15)$$

The remarkable thing here is that the constraint term $\boldsymbol{\theta} \boldsymbol{\theta}^T$ depends on $\boldsymbol{\theta}$ only, and so is constant for all sampling matrices \mathbf{B} , a great advantage when comparing different methods.

Since we are assuming flows are sampled independently, the Fisher information for N flows is just $N\mathbf{J}$, and the inverse becomes \mathcal{I}^+/N . For any unbiased estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, the CRLB then states that

$$E[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T] \geq \frac{\mathcal{I}^+(\boldsymbol{\theta})}{N}. \quad (16)$$

Because of independence we study $N = 1$. In practice all flows are sampled and so $N = N_f$.

Remark 6: There is an interpretation to the simple structure of the matrix $\mathbf{P} = \mathbf{J}^{-1} \mathbf{G} (\mathbf{G}^T \mathbf{J}^{-1} \mathbf{G})^{-1} \mathbf{G}^T \mathbf{J}^{-1}$. The scalar value $\mathbf{1}_W^T \mathcal{I}^+ \mathbf{1}_W$ is equivalent to $\text{Var}(\sum_{i=1}^W \hat{\theta}_i)$. By the equality constraint, we expect the best estimator of $\sum_{i=1}^W \hat{\theta}_i$ to have a variance of 0, since the estimator already knows that $\sum_{i=1}^W \theta_i = 1$. This corresponds to a CRLB of zero, namely $\mathbf{1}_W^T \mathcal{I}^+ \mathbf{1}_W = \mathbf{1}_W^T \mathbf{J}^{-1} \mathbf{1}_W - \mathbf{1}_W^T \boldsymbol{\theta} \boldsymbol{\theta}^T \mathbf{1}_W = \mathbf{1}_W^T \mathbf{diag}(\theta_1, \dots, \theta_W) \mathbf{1}_W - 1 = 0$. Thus, $\mathbf{P} = \boldsymbol{\theta} \boldsymbol{\theta}^T$ is the form of the correction term to the unconstrained covariance matrix \mathbf{J}^{-1} needed to satisfy the constraint.

III. THE SAMPLING METHODS

In this section we define the sampling methods we consider and derive their main properties. We begin with methods which have been described elsewhere, including simple packet and flow sampling, as well as others exploiting protocol information, in particular those proposed in [4], [1]. Apart from their inherent interest, we revisit these because in the unconditional framework these methods are now all **different** to before. More importantly, we also derive inverses analytically which has not been possible before, and thereby obtain a number of important insights. We also include the widely cited ‘Sample and Hold’ [6] whose Fisher information has not previously been studied. We then introduce our new method, Dual Sampling.

To better see the connection between the usual framework and ours, recall that b_{jk} is always a conditional probability with respect to the size k of the original flow. Typically however, it is also made conditional with respect to j , but we do not so here. Hence, if \mathbf{B}_c is the usual j -conditional matrix, then $\mathbf{B}_c \mathbf{C} = \tilde{\mathbf{B}}$ where $\mathbf{C} = \mathbf{I}_W - \mathbf{diag}(b_{01}, \dots, b_{0W})$, i.e. the matrix \mathbf{C}^{-1} does the conditioning.

We use the decomposition of (9) to describe each sampling matrix \mathbf{B} . In each case we define \mathbf{B} and $\tilde{\mathbf{B}}$, give the inverse $\tilde{\mathbf{B}}^{-1}$ of $\tilde{\mathbf{B}}$, and give explicit expressions for the diagonal terms $(\mathbf{J}^{-1})_{jj}$, or in some case for the entire inverse \mathbf{J}^{-1} . The importance of the diagonal terms will become very clear in Section IV.

A. Packet Sampling (PS)

By this we mean the simplest form of sampling, *i.i.d. packet sampling*, where each packet is retained independently with probability p_p and otherwise dropped with $q_p = 1 - p_p$. For the purpose of simplicity, we treat both ‘1 in N ’ periodic sampling and *i.i.d.* random sampling under the same framework, as both methods were shown to be statistically indistinguishable in practice [1].

The chief benefit of PS is its simplicity, and the fact that it can be implemented at high speed because a sampling decision can be made without even inspecting the packet. The chief disadvantage is the fact that it has a strong length bias, small flows are very likely to evaporate.

It is easy to see that $b_{jk} = \binom{k}{j} p_p^j q_p^{k-j}$, or

$$\mathbf{B} = \begin{bmatrix} q_p & q_p^2 & q_p^3 & q_p^4 & \cdots & q_p^W \\ p_p & 2p_p q_p & 3p_p q_p^2 & 4p_p q_p^3 & \cdots & \binom{W}{1} p_p q_p^{W-1} \\ 0 & p_p^2 & 3p_p^2 q_p & 6p_p^2 q_p^2 & \cdots & \binom{W}{2} p_p^2 q_p^{W-2} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & p_p^W \end{bmatrix}.$$

Before finding the inverse of $\tilde{\mathbf{B}}$, it is first instructive to note the following general results by Strum [13]. Let $\mathcal{B}(x, y)$ be an $(W+1) \times (W+1)$ matrix with the following structure

$$\mathcal{B}(x, y) = \begin{bmatrix} 1 & x & x^2 & \cdots & x^W \\ 0 & y & 2xy & \cdots & \binom{W}{1} x^{W-1} y \\ 0 & 0 & y^2 & \cdots & \binom{W}{2} x^{W-2} y^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & y^W \end{bmatrix}$$

which is known as the *binomial matrix*. Note that $\mathcal{B}(0, 1)$ reduces to the identity matrix. From [13] we have

Lemma 7: If $y \neq 0$, then $\mathcal{B}(x, y)$ is invertible and

$$[\mathcal{B}(x, y)]^{-1} = \mathcal{B}(-xy^{-1}, y^{-1}).$$

Using these results we can find the inverse of $\tilde{\mathbf{B}}$ (recall that $b'_{jk} = (\tilde{\mathbf{B}}^{-1})_{jk}$).

Theorem 8: The inverse of $\tilde{\mathbf{B}}$ is given by

$$\tilde{\mathbf{B}}^{-1} = \begin{bmatrix} p_p^{-1} & -2p_p^{-2}q_p & 3p_p^{-3}q_p^2 & \cdots & Wp_p^{-W}(-q_p)^{W-1} \\ 0 & p_p^{-2} & -3p_p^{-3}q_p & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \cdots & p_p^{-W} \end{bmatrix}$$

that is $b'_{jk} = (-1)^{k-j} \binom{k}{j} q_p^{k-j} p_p^{-k}$.

Proof: Here $\mathbf{b}_0^T = [q_p, q_p^2, \dots, q_p^W]$. We have

$$\mathcal{B}(q_p, p_p) = \begin{bmatrix} 1 & \mathbf{b}_0^T \\ \mathbf{0}_W & \tilde{\mathbf{B}} \end{bmatrix}.$$

Since $\mathcal{B}(q_p, p_p)[\mathcal{B}(q_p, p_p)]^{-1} = \mathbf{I}_{W+1}$, for some \mathbf{k} we have

$$\begin{bmatrix} 1 & \mathbf{b}_0^T \\ \mathbf{0}_W & \tilde{\mathbf{B}} \end{bmatrix} \begin{bmatrix} 1 & \mathbf{k} \\ \mathbf{0}_W & \tilde{\mathbf{B}}^{-1} \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{0}_W^T \\ \mathbf{0}_W & \mathbf{I}_W \end{bmatrix}.$$

Furthermore, from Lemma 7 we have $[\mathcal{B}(q_p, p_p)]^{-1} = \mathcal{B}(-q_p p_p^{-1}, p_p^{-1})$. Thus $\tilde{\mathbf{B}}^{-1}$ is essentially a principal $W \times W$ submatrix of $\mathcal{B}(-q_p p_p^{-1}, p_p^{-1})$. ■

For PS \mathbf{J}^{-1} is difficult to write in a compact form and will be omitted. It is however feasible to give using equation (12)

$$(\mathbf{J}^{-1})_{jj} = \sum_{k=j}^W \binom{k}{j}^2 q_p^{2(k-j)} p_p^{-2k} d_k - \frac{\left(\sum_{k=j}^W (-1)^{2k-j-1} d_k \binom{k}{j} q_p^{2k-j} p_p^{-2k} \right)^2}{\sum_{k=0}^W q_p^{2k} p_p^{-2k} d_k}. \quad (17)$$

The above form is derived in Appendix C.

B. Packet Sampling with Sequence Numbers (PS+SEQ)

First PS with parameter p_p is performed as above. Sequence numbers are then used as follows. Let s_l be the lowest sequence number among the sampled packets, and s_h the highest. All packets in-between these can now reliably inferred, hence $j = s_h - s_l + 1$ is the number of sampled packets returned. This is called ‘ALL-seq-sflag’ in [4].

The chief benefit of PS+SEQ is the fact that a potentially large number of packets can be ‘virtually’ observed without having to physically sample them. The disadvantage is the additional processing involved to perform the inference. Also, the technique is of limited value if flows are too short (as we discuss later).

If $j = 0, 1$ then sequence numbers cannot help and b_{jk} is as for PS. Otherwise, note that the j packets must occur in a contiguous block bordered by s_l and s_h . There are $k - j + 1$ possible positions for such a block, each characterized by $k - j$ unsampled packets outside it and the borders s_l and s_h . It follows that $b_{jk} = (k - j + 1) p_p^2 q_p^{k-j}$ for $2 \leq j \leq k$. Hence

$$\mathbf{B} = \begin{bmatrix} q_p & q_p^2 & q_p^3 & q_p^4 & \cdots & q_p^W \\ p_p & 2p_p q_p & 3p_p q_p^2 & 4p_p q_p^3 & \cdots & W p_p q_p^{W-1} \\ 0 & p_p^2 & 2p_p^2 q_p & 3p_p^2 q_p^2 & \cdots & (W-1) p_p^2 q_p^{W-2} \\ 0 & 0 & p_p^2 & 2p_p^2 q_p & \cdots & (W-2) p_p^2 q_p^{W-3} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & p_p^2 \end{bmatrix}.$$

Theorem 9: The inverse of $\tilde{\mathbf{B}}$ is

$$\tilde{\mathbf{B}}^{-1} = \begin{bmatrix} p_p^{-1} & -2q_p p_p^{-2} & q_p^2 p_p^{-2} & 0 & \cdots & 0 \\ 0 & p_p^{-2} & -2q_p p_p^{-2} & q_p^2 p_p^{-2} & \cdots & 0 \\ 0 & 0 & p_p^{-2} & -2q_p p_p^{-2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & p_p^{-2} \end{bmatrix}.$$

Proof: Observe that $\tilde{\mathbf{B}} = \mathbf{S}\mathbf{T}$ where

$$\mathbf{S} = \text{diag}(p_p, p_p^2, p_p^2, \dots, p_p^2)$$

is a $W \times W$ matrix and

$$\mathbf{T} = \begin{bmatrix} 1 & 2q_p & 3q_p^2 & 4q_p^3 & \cdots & Wq_p^{W-1} \\ 0 & 1 & 2q_p & 3q_p^2 & \cdots & (W-1)q_p^{W-2} \\ 0 & 0 & 1 & 2q_p & \cdots & (W-2)q_p^{W-3} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}.$$

A straightforward computation yields

$$\mathbf{T}^{-1} = \begin{bmatrix} 1 & -2q_p & q_p^2 & 0 & \cdots & 0 \\ 0 & 1 & -2q_p & q_p^2 & \cdots & 0 \\ 0 & 0 & 1 & -2q_p & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix},$$

and $\mathbf{S}^{-1} = \text{diag}(p_p^{-1}, p_p^{-2}, p_p^{-2}, \dots, p_p^{-2})$. Thus, $\tilde{\mathbf{B}}^{-1} = \mathbf{T}^{-1}\mathbf{S}^{-1}$, which proves our result. ■

The diagonal elements of \mathbf{J}^{-1} are given by

$$\begin{aligned} (\mathbf{J}^{-1})_{11} &= p_p^{-2}d_1 + 4q_p^2p_p^{-4}d_2 + q_p^4p_p^{-4}d_3 - \frac{(q_p^2p_p^{-4}d_1 + 2q_p^3p_p^{-4}d_2)^2}{r} \\ (\mathbf{J}^{-1})_{22} &= p_p^{-4}d_2 + 4q_p^2p_p^{-4}d_3 + q_p^4p_p^{-4}d_4 - \frac{q_p^4p_p^{-8}d_2^2}{r} \\ (\mathbf{J}^{-1})_{jj} &= p_p^{-4}d_j + 4q_p^2p_p^{-4}d_{j+1} + q_p^4p_p^{-4}d_{j+2}, \quad 3 \leq j \leq W \end{aligned}$$

where $r = d_0 + q_p^2p_p^{-2}d_1 + q_p^4p_p^{-4}d_2$, and for convenience we set $d_j = 0$ for $j > W$.

C. Packet Sampling with SYN Sampling (PS+SYN)

First PS with parameter p_p is performed as above. A post-processing phase then discards all packets belonging to sampled flows which lack a SYN packet (or more accurately, maps them to sampled flows with $j = 0$). This was introduced in [1] and called ‘‘SYN-pktct’’ in [4].

The chief benefit of PS+SYN is that the flow length bias of PS is averted by keeping flows based on the presence of the SYN, which is flow length independent. The chief disadvantage is the fact that it is wasteful: if $p_p = 0.01$ then 99% of packets which were initially sampled belong to ‘failed’ flows and are subsequently discarded!

A flow evaporates iff its SYN is not sampled, hence $b_{0k} = q_p$. For $j \geq 1$ the SYN must first be sampled, which occurs with probability p_p , and conditional on this $j-1$ more packets must be sampled from the remaining $k-1$ using i.i.d. sampling. Hence $b_{jk} = p_p \cdot \binom{k-1}{j-1} p_p^{j-1} q_p^{k-j}$ for $j \geq 1$, giving

$$\mathbf{B} = \begin{bmatrix} q_p & q_p & q_p & \cdots & q_p \\ p_p & p_p q_p & p_p q_p^2 & \cdots & p_p q_p^{W-1} \\ 0 & p_p^2 & 2p_p^2 q_p & \cdots & (W-1)p_p^2 q_p^{W-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & p_p^W \end{bmatrix}.$$

Theorem 10: The inverse of $\tilde{\mathbf{B}}$ is given by

$$\tilde{\mathbf{B}}^{-1} = \frac{1}{p_p} \begin{bmatrix} 1 & -q_p p_p^{-1} & q_p^2 p_p^{-2} & \cdots & (-q_p)^{W-1} p_p^{-W+1} \\ 0 & p_p^{-1} & -2q_p p_p^{-2} & \cdots & (-q_p)^{W-2} p_p^{-W+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \cdots & p_p^{-W+1} \end{bmatrix}$$

that is $b'_{jk} = (-1)^{k-j} \binom{k-1}{j-1} q_p^{k-j} p_p^{-k}$.

Proof: Note that we can express $\tilde{\mathbf{B}}$ in terms of a $W \times W$ binomial matrix $\mathcal{B}(x, y)$ such that $\tilde{\mathbf{B}} = p_p \mathcal{B}(q_p, p_p)$. Then by Lemma 7, the inverse is given by $\tilde{\mathbf{B}} = (1/p_p) \mathcal{B}(-q_p p_p^{-1}, p_p^{-1})$. ■

It is easy to see that the condition of Corollary 3 is satisfied with $p = p_p$. Hence $\mathbf{J}^{-1} = \tilde{\mathbf{J}}^{-1} - \frac{q_p}{p_p} \boldsymbol{\theta} \boldsymbol{\theta}^T$, and the diagonal entries for $1 \leq j \leq W$ are

$$(\mathbf{J}^{-1})_{jj} = \sum_{k=j}^W \binom{k-1}{j-1}^2 q_p^{2(k-j)} p_p^{-2k} d_k - \frac{q_p}{p_p} \theta_j^2. \quad (18)$$

D. Packet Sampling with SYN and SEQ (PS+SYN+SEQ)

First sampling is performed according to PS+SYN with parameter p_p , and then on each resulting sampled flow the sequence number post-processing is performed as per PS+SEQ. This is called ‘‘SYN-seq’’ in [4]. PS+SYN+SEQ is a hybrid of PS+SYN and PS+SEQ and combines the advantages and disadvantages of both.

If $j = 0, 1$ then sequence numbers cannot help and b_{jk} is as for PS+SYN. Otherwise, by combining the arguments above, it is easy to see that $b_{jk} = p_p \cdot p_p q_p^{k-j}$ for $j > 1$, giving

$$\mathbf{B} = \begin{bmatrix} q_p & q_p & q_p & q_p & \cdots & q_p \\ p_p & p_p q_p & p_p q_p^2 & p_p q_p^3 & \cdots & p_p q_p^{W-1} \\ 0 & p_p^2 & p_p^2 q_p & p_p^2 q_p^2 & \cdots & p_p^2 q_p^{W-2} \\ 0 & 0 & p_p^2 & p_p^2 q_p & \cdots & p_p^2 q_p^{W-3} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & p_p^2 \end{bmatrix}. \quad (19)$$

Theorem 11: The inverse of $\tilde{\mathbf{B}}$ is given by

$$\tilde{\mathbf{B}}^{-1} = \frac{1}{p_p} \begin{bmatrix} 1 & -\frac{q_p}{p_p} & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{p_p} & -\frac{q_p}{p_p} & 0 & \cdots & 0 \\ 0 & 0 & \frac{1}{p_p} & -\frac{q_p}{p_p} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & \frac{1}{p_p} \end{bmatrix}$$

Proof: A straightforward computation shows that $\tilde{\mathbf{B}}\tilde{\mathbf{B}}^{-1} = \mathbf{I}_W$. ■

Since $\mathbf{b}_0 = q_p \mathbf{1}_W$, Corollary 3 applies and states that $\mathbf{J}^{-1} = \tilde{\mathbf{J}}^{-1} - \frac{q_p}{p_p} \boldsymbol{\theta} \boldsymbol{\theta}^T$. The diagonal elements can be explicitly written, but we defer this to Section III-H.

E. Flow Sampling (FS)

In i.i.d. flow sampling [3], flows are retained independently with probability p_f and otherwise dropped with $q_f = 1 - p_f$.

The chief benefit of FS is the fact that flows which are sampled retain their full complement of packets, eliminating completely the difficulties in inverting sampled flow sizes back to original sizes. The chief disadvantage is that each packet requires a lookup in a flow table to see if it belongs to flow which has been sampled.

A flow evaporates iff its SYN is not sampled, hence $b_{0k} = q_f$. If a flow has been selected, which occurs with probability

p_f , then conditional on this $j = k$ with certainty, that is $b_{jk} = 1$ if $j = k$, else 0, for $j \geq 1$:

$$\mathbf{B} = \begin{bmatrix} q_f & q_f & q_f & q_f & \cdots & q_f \\ p_f & 0 & 0 & 0 & \cdots & 0 \\ 0 & p_f & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & p_f \end{bmatrix}. \quad (20)$$

The inverse of $\tilde{\mathbf{B}}$ is just $\tilde{\mathbf{B}}^{-1} = \mathbf{I}_W/p_f$, and \mathbf{J} takes the elegant form $\mathbf{J}(\boldsymbol{\theta}) = q_f \mathbf{1}_W \mathbf{1}_W^T + p_f \text{diag}(\theta_1^{-1}, \dots, \theta_W^{-1})$.

Clearly Corollary 3 applies with $p = p_f$. Since

$$\tilde{\mathbf{J}} = p_f \text{diag}(\theta_1^{-1}, \dots, \theta_W^{-1}),$$

the unconstrained inverse can therefore be expressed as

$$\mathbf{J}^{-1}(\boldsymbol{\theta}) = \frac{1}{p_f} \text{diag}(\boldsymbol{\theta}) + (1 - \frac{1}{p_f}) \boldsymbol{\theta} \boldsymbol{\theta}^T. \quad (21)$$

Remark 12: By using equations (15) and (21), the inverse constrained Fisher information matrix for FS is given by

$$\mathcal{I}^+ = \frac{1}{p_f} \text{diag}(\boldsymbol{\theta}) - \frac{1}{p_f} \boldsymbol{\theta} \boldsymbol{\theta}^T,$$

with the diagonals being $(\mathcal{I}^+)_{kk} = (1/p_f)\theta_k(1 - \theta_k)$. This is just an appropriately scaled version (by $1/p_f$) of the inverse Fisher information of a multinomial model. This makes sense given that FS works by picking out whole flows from the original flow set in an i.i.d fashion and that the complete likelihood function can be modeled using a multinomial model.

F. Packet Sampling with SYN, FIN, SEQ (PS+SYN+FIN+SEQ)

In this scheme SYN packets are retained independently with probability p_f and otherwise dropped with $q_f = 1 - p_f$, and the FIN packets corresponding to sampled SYNs are also sampled, but no others. Sequence numbers are then used to infer flow sizes.

This scheme has two great advantages: like FS the flows sampled are sampled perfectly, and moreover this could be achieved by physically sampling only two packets per flow, based on looking for SYN and FIN flags on a per packet basis, which is feasible at high speed. The disadvantage is that a moderate minority of flows do not terminate correctly with a FIN, and/or the FIN may be not observable. Furthermore, flows consisting of a single SYN (such as in a SYN attack) would be entirely missed. For this reason we choose not to study it further.

Information theoretically, PS+SYN+FIN+SEQ is identical to flow sampling provided we assume $\theta_1 \approx 0$.

G. Sample and Hold (SH)

Here packets are first sampled as for PS with probability p_p , however for each flow if a packet is sampled, then all subsequent packets in the flow will be. Hence the total number of packets sampled is much higher than the parameter p_p . The scheme was introduced in [6].

The chief benefit of SH is that provided just a single packet from a flow is PS-sampled, then typically many will be finally

sampled. This conditional behaviour is much more effective than methods using SEQ where at least two packets must be PS-sampled, and even then fewer packets are finally recouped. Essentially SH skips a geometric number of the first packets in a flow and then captures all the rest. It therefore efficiently skips small flows and accurately recovers the size of large flows. This amplified flow length bias (even stronger than for PS) makes it well suited for the heavy hitter problem (i.e. accurately measuring the very largest flows) for which it was originally designed. For flow size estimation more generally however, it is a disadvantage for most flow sizes. The other disadvantage is the need to check, for **each packet**, whether it belongs to a flow which has already been sampled. This makes it very costly in a true sampling implementation. Indeed Estan and Varghese implemented it using lossy sketching techniques [6].

A flow evaporates iff none of its packets are sampled, hence $b_{0k} = q_p^k$. Otherwise, $b_{jk} = p_p q_p^{k-j}$, thus:

$$\mathbf{B} = \begin{bmatrix} q_p & q_p^2 & q_p^3 & q_p^4 & \cdots & q_p^W \\ p_p & p_p q_p & p_p q_p^2 & p_p q_p^3 & \cdots & p_p q_p^{W-1} \\ 0 & p_p & p_p q_p & p_p q_p^2 & \cdots & p_p q_p^{W-2} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & p_p \end{bmatrix}. \quad (22)$$

It is interesting to note that the first row is as for PS, the second as for PS+SYN, and subsequent rows like PS+SYN scaled by $1/p_p$.

Theorem 13: The inverse of $\tilde{\mathbf{B}}$ is given by

$$\tilde{\mathbf{B}}^{-1} = \frac{1}{p_p} \begin{bmatrix} 1 & -q_p & 0 & 0 & \cdots & 0 \\ 0 & 1 & -q_p & 0 & \cdots & 0 \\ 0 & 0 & 1 & -q_p & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -q_p \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}$$

Proof: It is easily verified that $\tilde{\mathbf{B}}\tilde{\mathbf{B}}^{-1} = \mathbf{I}_W$. ■

It is not difficult to show that Proposition 2 reduces to $\mathbf{J}^{-1} = \tilde{\mathbf{J}}^{-1} - \mathbf{C}$, where the only non-zero element of \mathbf{C} is $\mathbf{C}_{11} = p_p^{-2} q_p^2 d_1^2 (p_p^2 d_0 + q_p^2 d_1)^{-1} = d_0/p_p$ since $d_1 = \frac{p_p}{q_p} d_0$. Furthermore, using (11) one can show that $\tilde{\mathbf{J}}^{-1}$ is tridiagonal (and symmetric) with upper off-diagonal terms $(\tilde{\mathbf{J}}^{-1})_{k,k+1} = -p_p^{-2} q_p d_{k+1}$, $k < W$, and diagonal elements

$$\begin{aligned} (\mathbf{J}^{-1})_{11} &= \frac{1}{p_p^2} (d_1 + q_p^2 d_2) - \frac{1}{p_p} d_0 \\ &= \theta_1 + \frac{1}{p_p} \sum_{k=2}^W q_p^{k-1} \theta_k \end{aligned} \quad (23)$$

$$\begin{aligned} (\mathbf{J}^{-1})_{jj} &= \frac{1}{p_p^2} (d_j + q_p^2 d_{j+1}), \quad 2 \leq j \leq W-1 \\ &= \frac{\theta_j}{p_p} + \frac{1+q_p}{p_p} \sum_{k=j+1}^W q_p^{k-j} \theta_k \end{aligned} \quad (24)$$

$$(\mathbf{J}^{-1})_{WW} = \frac{d_W}{p_p^2} = \frac{\theta_W}{p_p},$$

where we have used the property of \mathbf{B} that $d_j = p_p \theta_j + q_p d_{j+1}$ for $1 \leq j \leq W - 1$.

H. Dual Sampling (DS)

DS can be defined simply as follows. First, at the packet level it consists of two PS schemes running in parallel, one which operates only on SYN packets with sampling probability p_f , and the other only on non-SYN packets with sampling probability p_p . In a second phase, sampled flows which lack a SYN are discarded, and sequence numbers are used to infer additional ‘virtual’ packets, as in PS+SYN+SEQ. Thus, at one level DS is simply a generalization of PS+SYN+SEQ, and reduces to it when $p_f = p_p$. However, the generalization is significant as it also includes FS as the special case $p_p = 1$, and interpolates continuously between the two. This is illustrated in Figure 2 which depicts the (p_p, p_f) parameter space, and marks the special cases.

Dual Sampling is ‘dual’ in two senses. Computationally it can be viewed as the original PS sampling being split into two, at the (low) cost of per-packet switching based on some bit checking to determine which PS applies. Information theoretically, the sampling is now split into two parts with very different natures, each controlled by a dedicated parameter: a FS-like direct sampling of flows, and a PS-like in-flow sampling. Here p_f controls the number of sampled flows, and p_p their ‘quality’.

The derivation of the sampling matrix mirrors closely that of PS+SYN+SEQ. The result shows clearly how p_f and p_p act in a modular fashion. The flow sampling component controls the top row of \mathbf{B} and factors $\tilde{\mathbf{B}}$, whereas the packet sampling component determines the internal structure of $\tilde{\mathbf{B}}$.

$$\mathbf{B} = \begin{bmatrix} q_f & q_f & q_f & q_f & \cdots & q_f \\ p_f & p_f q_p & p_f q_p^2 & p_f q_p^3 & \cdots & p_f q_p^{W-1} \\ 0 & p_f p_p & p_f p_p q_p & p_f p_p q_p^2 & \cdots & p_f p_p q_p^{W-2} \\ 0 & 0 & p_f p_p & p_f p_p q_p & \cdots & p_f p_p q_p^{W-3} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & p_f p_p \end{bmatrix}.$$

The separation of the FS and PS roles in $\tilde{\mathbf{B}}$ is clearly reflected in its inverse:

$$\tilde{\mathbf{B}}^{-1} = \frac{1}{p_f} \begin{bmatrix} 1 & -\frac{q_p}{p_p} & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{p_p} & -\frac{q_p}{p_p} & 0 & \cdots & 0 \\ 0 & 0 & \frac{1}{p_p} & -\frac{q_p}{p_p} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & \frac{1}{p_p} \end{bmatrix}.$$

The similarity between $\tilde{\mathbf{B}}^{-1}$ for SH and DS is striking. Indeed, whereas SH uses sampling to select which flows it will focus on and then holds to them, DS reverses these operations and thus could be described as a ‘Hold and Sample’ scheme. However, although they each combine PS and FS features, the methods remain significantly different. In particular SH is strongly flow length biased whereas DS is unbiased.

Once again Corollary 3 for the inverse Fisher information holds, showing that $\mathbf{J}^{-1} = \tilde{\mathbf{J}}^{-1} - \frac{q_f}{p_f} \boldsymbol{\theta} \boldsymbol{\theta}^T$. Similarly to SH,

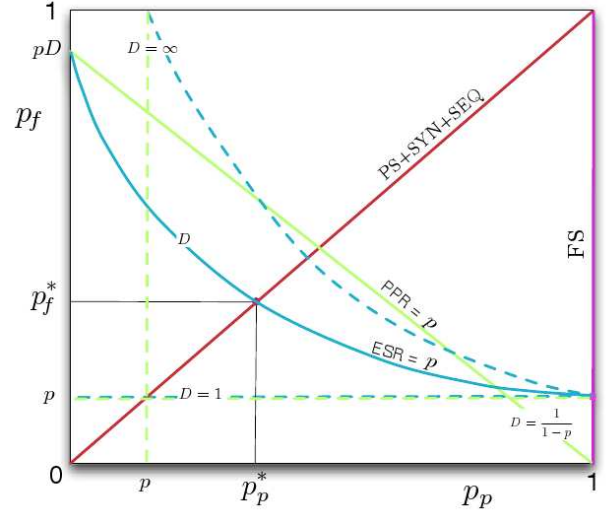


Fig. 2. Parameter space (p_p, p_f) of DS. Fixing $\text{ESR} = p$ constrains the family to the solid (blue) curve $p_f(p_p; p)$, where it reduces to PS+SYN+SEQ at $p_p = p_p^*$ and FS at $p_p = 1$. Fixing $\text{PPR} = p$ constrains the family to a straight (green) line emanating from $(0, pD)$. For fixed p , the constraint curves depend on D . Three examples are given for each normalization.

from (11) one can show that $\tilde{\mathbf{J}}^{-1}$ is tridiagonal with upper off diagonal terms $(\tilde{\mathbf{J}}^{-1})_{k,k+1} = -(p_f p_p)^{-2} q_p d_k$, $k < W$. The diagonal elements (here $2 \leq j \leq W - 1$) are given by

$$\begin{aligned} (\mathbf{J}^{-1})_{11} &= \frac{1}{p_f^2 p_p^2} (p_p^2 d_1 + q_p^2 d_2) - \frac{q_f}{p_f} \theta_1^2 \\ &= \frac{\theta_1}{p_f} + \frac{1}{p_f p_p} \sum_{k=2}^W q_p^{k-1} \theta_k - \frac{q_f}{p_f} \theta_1^2 \end{aligned} \quad (25)$$

$$\begin{aligned} (\mathbf{J}^{-1})_{jj} &= \frac{1}{p_f^2 p_p^2} (d_j + q_p^2 d_{j+1}) - \frac{q_f}{p_f} \theta_j^2 \\ &= \frac{\theta_j}{p_f p_p} + \frac{1}{p_f p_p} \sum_{k=j+1}^W q_p^{k-j} (1 + q_p) \theta_k - \frac{q_f}{p_f} \theta_j^2 \\ (\mathbf{J}^{-1})_{WW} &= \frac{d_W}{p_f^2 p_p^2} - \frac{q_f}{p_f} \theta_W^2 = \frac{\theta_W}{p_f p_p} - \frac{q_f}{p_f} \theta_W^2. \end{aligned} \quad (26)$$

By setting $p_f = p_p$ we obtain those for PS+SYN+SEQ.

IV. COMPARISONS

In this section we compare and contrast the performance of the different methods, using two normalizations which are the key to a fair comparison. We show that a positive semidefinite comparison holds for certain cases, and justify why we ultimately resort to comparison of the diagonals of the covariance matrix. We provide a partial ranking of the methods. Proofs of most key results in this section are deferred to Appendix D.

A. Normalization

We must first consider how to compare fairly. We do this in two ways, using the following packet-based measures of incoming workload/information.

- **Packet Processing Rate (PPR)**: the rate at which packets are initially being sampled (and hence require some further processing).
- **Effective Sampling Rate (ESR)**: the rate at which packets are arriving to the flow table (and hence become available as information for estimation).

PPR is a measure of the processing speed required by the methods, whereas ESR is a measure of the arrival rate of packets containing information which is actually used by the method. Because the methods which discard flows without SYN packets lose (the majority) of their packets, an equal PPR comparison greatly disadvantages them. An equal ESR comparison effectively inflates the parameters of such methods to compensate.

Note that given a sampling method X , both the PPR and ESR normalizations of X +SEQ and X are identical, as the methods only differ in how the packet data is used, not how many packets are physically collected. We now present the normalizations for each method. Here and below we use $D = \sum_{k=1}^W k\theta_k \geq 1$ to denote the average flow size.

Table I gives the results for PPR. Both the average sampling rate p as a function of the method parameter(s), and its inverse, the parameter value required to set the average sampling rate at p , are shown. The expression $p(p_p, \theta)$ for SH was derived by computing the average number of packets sampled (which is highly dependent on θ) and dividing the result by the average flow size D . It is monotonically increasing in p_p with $p(1, \theta) = 1$, and $p/p_p \rightarrow \sum_{j=1}^W j \sum_{k=j}^W \theta_k / D \geq 1$ as $p_p \rightarrow 0$, which in the worst case ($\theta_W = 1$) becomes $p/p_p = (1+W)/2$. We invert $p(p_p, \theta)$ numerically as required for choices of θ , denoted by $\text{root}(p(p_p))$ in the tables. DS has two parameters. We choose to make p_p the independent one.

Table II gives the results for ESR. Only the methods which involve discarding packets after their initial collection have changed compared to Table I. Note that for PS+SYN $p/p_p \rightarrow 1/D \geq 1$ as $p_p \rightarrow 0$, and $p(1) = 1$.

We are particularly interested in the ESR case for the DS family, and so repeat its ESR normalization here:

$$p_f(p_p; p) = \frac{pD}{p_p(D-1) + 1}. \quad (27)$$

The denominator $p_p(D-1) + 1$ is simply the average sampled flow size, conditional on its SYN being sampled. For fixed p , this equation gives p_f as a monotonically decreasing, in fact convex, function of p_p . Three examples (the blue curves) are given in Figure 2 for different values of D . To maintain an ESR fixed at p , if we increase p_p we must move down the curve and decrease p_f to compensate. The PPR case is similar

Method(params)	Average sampling rate $p =$	PPR(p) =
PS(p_p)	p_p	p
PS+SYN(p_p)	p_p	p
FS(p_f)	p_f	p
SH(p_p)	$\frac{p_p}{D} \sum_{j=1}^W j q_p^{-j} \sum_{k=j}^W q_p^k \theta_k$	$\text{root}(p(p_p))$
DS(p_p, p_f)	$p_f \left(\frac{1}{D} \right) + p_p \left(1 - \frac{1}{D} \right)$	$p_f = pD - p_p(D-1)$

TABLE I
AVERAGE PPR SAMPLING RATES AND NORMALIZATIONS.

Method(params)	Average sampling rate $p =$	ESR(p) =
PS(p_p)	p_p	p
PS+SYN(p_p)	$\frac{p_p(p_p(D-1)+1)}{D}$	$\frac{-1+\sqrt{1+4pD(D-1)}}{2(D-1)}$
FS(p_f)	p_f	p
SH(p_p)	$\frac{p_p}{D} \sum_{j=1}^W j q_p^{-j} \sum_{k=j}^W q_p^k \theta_k$	$\text{root}(p(p_p))$
DS(p_p, p_f)	$\frac{p_f(p_p(D-1)+1)}{D}$	$p_f = \frac{pD}{p_p(D-1)+1}$

TABLE II
AVERAGE ESR SAMPLING RATES AND NORMALIZATIONS.

but simpler as the level curves are straight lines (green lines in Figure 2). Note that depending on p and D , under both PPR and ESR there are values of p_f and/or p_p that may be disallowed.

To compare methods we also require a performance metric. A natural criterion is the set of diagonal elements of the CRLB, the k -th being

$$(\mathcal{I}^+)_{kk} = (\mathbf{J}^{-1})_{kk} - \theta_k^2, \quad (28)$$

since this is a lower bound on the variance of any unbiased estimator of θ_k . We plot the square root of these values, calculated using the expressions for the previous section.

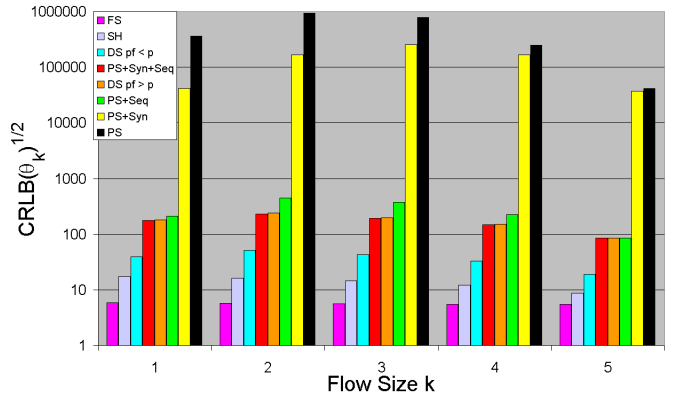


Fig. 3. An equal PPR comparison of the CRLB bound for θ with $p = 0.005$.

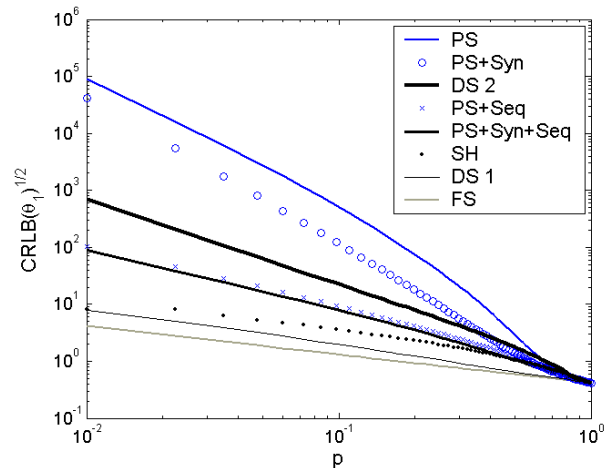


Fig. 4. Dependence of the CLRb bound on p , PPR comparison.

We begin with some instructive examples. Consider the results of Figure 3, where we set $W = 5$ with

$$\theta = \{0.22, 0.21, 0.20, 0.19, 0.08\},$$

for which $D = 2.90$ packets. We use the PPR normalization with $p = 0.005$, corresponding to $p_p \approx 0.0021845$ for SH. For DS we give two examples satisfying p : $(p_f, p_p) = (0.001, 0.0443)$, and $(p_f, p_p) = (0.1, 0.00334)$.

As expected from earlier work, the performance of PS is extraordinarily poor. In agreement with the results of [4] and as expected, the inclusion of SEQ improves it enormously, by orders of magnitude, but it is still orders of magnitude behind FS, which has the lowest standard deviation bound of all. In a highly counterintuitive result, PS+SYN performs (much!) better than PS, despite the enormous number of wasted packets. We can offer two reasons for this. First, in the unconditional framework information in discarded flows is not entirely wasted, some is recouped by the (observable) increase in the $j = 0$ outcome¹. Second, SYN sampled flows (much like FS) have no flow length bias. Together these lead PS+SYN to perform better in most cases even under PPR. Three results for DS are given. From best to worst these were with $p_f = 0.001 < p$, the special case PS+SYN+SEQ with $p_f = p$, and $p_f = 0.1 > p$. Finally SH performs well in this example as its key disadvantage, a strong bias to large flows, does not play a large role for such a small value of W .

Using the same PPR based comparison, Figure 4 shows the improvement in CRLB as p increases, as we expect. We see that very high values of p are needed before the performance of FS is approached. Here the free parameter p_f for DS was chosen to guarantee meaningful examples to either side of PS+SYN+SEQ. Specifically, the normalization defines for each value of p a feasible range $[p_f^l, p_f^u]$ for p_f which includes $p_f = p$. For DS1 we set $p_f = p - c(p - p_f^l) < p$, and for DS2 $p_f = p + c(p_f^u - p) > p$, where $c \in [0, 1]$. In the figure we use $c = 0.1$.

The great merit of the ESR normalization is that it allows us to view the methods as being different ways in which the

¹It was pointed out in [4], under a conditional framework, that PS+SYN+SEQ outperforms PS+SEQ on actual traces. Our own numerical evaluation of the CRLB of these methods (under the conditional framework) also show this in some cases, leading us to conjecture the same counterintuitive results may hold for the conditional framework.

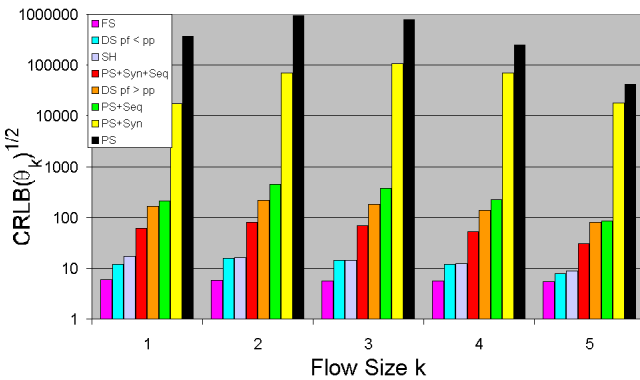


Fig. 5. An equal ESR comparison of the CRLB bound for θ with $p = 0.005$.

same budget of sampled packets can be allocated among flows. The essential tradeoff is that we can have more sampled flows of poor quality, or fewer of higher quality. In a class with no flow length bias, flow sampling is at one extreme: for a given $\text{ESR} = p$ it gives the minimum number of sampled flows, each of perfect quality. Among sub-classes of methods with roughly the same number of flows, we can then distinguish between the finer details of how the ‘holes’ appear over the flow, which will have an impact on the degree of improvement which the sequence numbers can bring.

Results using the same scenario as Figure 3 but ESR normalized are shown in Figure 5. As expected, all SYN-based methods improve significantly. In particular the DS with the smaller $p_f = 0.001$ now outperforms SH, and is second only to FS.

Another example with a larger $W = 50$ and a truncated exponential distribution for θ with $D = 16.039$ is given in Figure 6. The same general conclusions hold, with the exception of SH, whose performance is now worse than PS+SYN+SEQ rather than considerably better. This is to be expected at larger W , as SH expends its packet budget on the rare largest flows. At realistic W values in network data this effect is further exaggerated.

It is interesting to note in the right plot of Figure 6 that variance decreases as k increases. This is consistent with the fact that the ambiguity inherent in an observation of j sampled packets is lower for larger j and disappears at $j = W$, since this observation can only arise in one way. The dip at very small k we attribute to the influence of the constraint.

B. Positive Semidefinite Comparisons

The examples above compared methods using the CRLB of each θ_k separately. Let two methods have Fisher information \mathbf{J}_1 and \mathbf{J}_2 . A more complete, and in a sense ideal comparison, would be to show that $\mathbf{J}_1 \geq \mathbf{J}_2$, since that would imply $\mathcal{I}_1^+ \leq \mathcal{I}_2^+$, a lower CRLB. What this positive semidefinite comparison really means is that for any linear combination $f(\theta) = \mathbf{a}^T \theta = \sum_k a_k \theta_k$ of the parameters, the (bound on the) variance of $f(\theta)$ under method 1 will be less than that under method 2. A geometric interpretation is that the ellipsoid corresponding to unit variance of vectors under \mathbf{J}_1 lies entirely within that of \mathbf{J}_2 . In this section we provide a number of comparisons of this type between methods, and explain why it is not suitable as a universal basis of comparison.

We first confirm the intuition that methods lose information monotonically in their sampling parameter(s).

Theorem 14: For any method Z surveyed here if $p_1 \geq p_2$ (for DS $p_{p,1} \geq p_{p,2}$ and $p_{f,1} \geq p_{f,2}$) then $\mathbf{J}_Z(p_1) \geq \mathbf{J}_Z(p_2)$ (for DS $\mathbf{J}_Z(p_{p,1}, p_{f,1}) \geq \mathbf{J}_Z(p_{p,2}, p_{f,2})$). Equality holds iff $p_1 = p_2$ (for DS $(p_{p,1}, p_{f,1}) = (p_{p,2}, p_{f,2})$).

The proof (see Appendix D), is a straightforward consequence of closure under sampling for all methods except SH.

Next, we confirm that the use of sequence numbers increases Fisher information.

Theorem 15: $\mathbf{J}_{Z+\text{SEQ}} \geq \mathbf{J}_Z$ for each of $Z = \text{PS}$ and $Z = \text{PS}+\text{SYN}$.

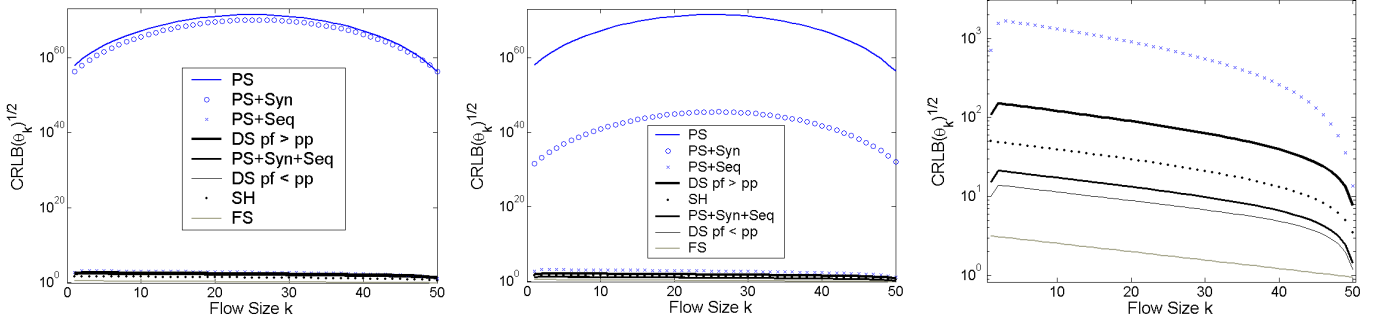


Fig. 6. Comparison of the CRLB bound with $W = 50$. Left: PPR with $p = 0.005$, $p_f = 0.001$, $p_p = 0.0052$ for DS ($p_f < p_p$) and $p_f = 0.01$, $p_p = 0.0047$ for DS ($p_f > p_p$). Middle: ESR with $p = 0.005$, $p_f = 0.032$, $p_p = 0.1$ for DS ($p_f < p_p$) and $p_f = 0.079$, $p_p = 0.001$ for DS ($p_f > p_p$). Right: Zoom of middle plot, the same legend applies.

Proof: We make use of the data processing inequality (DPI) for Fisher information (see Theorem 31 in Appendix B), which states that if $\theta \rightarrow Y \rightarrow X$ is a Markov chain, where X is a deterministic function of Y , then $\mathbf{J}_Y(\theta) \geq \mathbf{J}_X(\theta)$ which holds with equality if X is a sufficient statistic of Y . We first consider PS+SEQ and PS. Flows are selected randomly and are represented as a random vector of SEQ numbers $\mathbf{V} = \{A, A+1, \dots, A+K-1\}$ with realization $\mathbf{v} = \{a, a+1, \dots, a+K-1\}$. Here, K represents the flow size (realization k) while the A , the initial SEQ number (realization a) is uniform over $[0, N_a - 1]$. All operations on them are modulo N_a , thus allowing wrap-arounds. A is independent of K . We also assume that $W \leq N_a$ to avoid problems with multiple wrap-arounds.

After the PS process, we have the SEQ vector $\mathbf{Y} = \{Y_i, 1 \leq i \leq J\}$, (realization $\mathbf{y} = \{y_i, 1 \leq i \leq j\}$). We define two statistics on \mathbf{Y} : $J(\mathbf{Y})$, the actual number of raw packets sampled, and $L(\mathbf{Y}) = \max_i(Y_i) - \min_i(Y_i) + 1$ (with a sample $l(\mathbf{y})$ and $L(\mathbf{Y}) = 0$ if \mathbf{Y} is empty), the inferred length of the sequence. The statistic $L(\mathbf{Y})$ disregards A since it takes the difference of SEQ numbers. Note that both statistics are deterministic functions of \mathbf{Y} and so form Markov chains $\theta \rightarrow \mathbf{Y} \rightarrow L(\mathbf{Y})$ and $\theta \rightarrow \mathbf{Y} \rightarrow J(\mathbf{Y})$. A straightforward application of DPI yields $\mathbf{J}_Y \geq \mathbf{J}_{L(\mathbf{Y})}$ and $\mathbf{J}_Y \geq \mathbf{J}_{J(\mathbf{Y})}$.

We show that $L(\mathbf{Y})$ is a sufficient statistic of \mathbf{Y} w.r.t. θ . We use the Fisher-Neyman factorization theorem [14, Theorem 6.5, p. 35] which states that if the probability mass function of \mathbf{Y} takes the form

$$\Pr(\mathbf{Y} = \mathbf{y}) = h(\mathbf{y})g(l(\mathbf{y}), \theta), \quad (29)$$

where h is independent of the parameters θ , then $L(\mathbf{Y})$ is a sufficient statistic. Let $m \geq 0$ denote the number of unsampled packets before the first sampled packet. If $J = j > 0$,

$$\begin{aligned} \Pr(\mathbf{Y} = \mathbf{y}) &= \sum_{k=l}^W \theta_k \sum_{m=0}^{k-l} \Pr(A = y_1 - m) q_p^m p_p (p_p^{j-2} q_p^{l-j}) p_p q_p^{k-l-m} \\ &= \Pr(A = a) \sum_{k=l}^W \theta_k \sum_{m=0}^{k-l} q_p^{k-j} p_p^j \\ &= \frac{1}{N_a} p_p^j q_p^{-j} \left(\sum_{k=l}^W (k-l+1) q_p^k \theta_k \right), \end{aligned}$$

since A is uniform. For the case $j = 0$,

$$\Pr(\mathbf{Y} = \emptyset) = \sum_{k=1}^W q_p^k \theta_k.$$

Clearly, each case satisfies (29), hence $L(\mathbf{Y})$ is a sufficient statistic of θ .

By the DPI, since $L(\mathbf{Y})$ is a sufficient statistic, $\mathbf{J}_Y = \mathbf{J}_{L(\mathbf{Y})}$. From the previous relation $\mathbf{J}_Y \geq \mathbf{J}_{J(\mathbf{Y})}$, we now have $\mathbf{J}_{L(\mathbf{Y})} \geq \mathbf{J}_{J(\mathbf{Y})}$. But $J(\mathbf{Y})$ is equivalent to the statistic used in PS, proving the result.

As for PS+SYN sampling, we define an additional random variable S taking value 1 if the SYN packet was sampled, 0 otherwise. \mathbf{V} is defined as before. $\mathbf{Y} = \emptyset$ if and only if $S = 0$. The same statistics $L(\mathbf{Y})$ and $J(\mathbf{Y})$ are defined. Then, for $J = j > 0$,

$$\begin{aligned} \Pr(\mathbf{Y} = \mathbf{y}) &= \sum_{k=l}^W \theta_k \Pr(A = y_1) p_p (p_p^{j-2} q_p^{l-j}) p_p q_p^{k-l} \\ &= \Pr(A = a) \sum_{k=l}^W \theta_k q_p^{k-j} p_p^j \\ &= \frac{1}{N_a} p_p^j q_p^{-j} \left(\sum_{k=l}^W q_p^k \theta_k \right), \end{aligned}$$

and for $j = 0$,

$$\Pr(\mathbf{Y} = \emptyset) = \sum_{k=1}^W q_p \theta_k = q_p.$$

Once again, each case satisfies (29), hence $L(\mathbf{Y})$ is a sufficient statistic. Thus, by DPI, the same relationship $\mathbf{J}_{L(\mathbf{Y})} \geq \mathbf{J}_{J(\mathbf{Y})}$ holds, with $J(\mathbf{Y})$ now equivalent to the statistic used in PS+SYN. ■

Intuitively this result seems obvious: if the additional information afforded by the sequence numbers is available, we should certainly be able to do better by using it. However, it is tempting to conclude that by the same logic $\mathbf{J}_{PS} > \mathbf{J}_{PS+SYN}$ under the PPR normalization, since deciding to discard flows without a SYN is also a deterministic transformation. However, the data processing inequality does not apply here because some of the information used by PS+SYN (namely the SYN variable S), although available to PS, is not used by

it. Indeed we saw in the figures above that, counterintuitively, PS+SYN can actually outperform PS in terms of the individual variances under PPR, which is a counter-example to the more general positive semidefinite comparison. Under ESR this is even more true as we saw from the middle plot in Figure 6.

We now give another, even more surprising counter-example which teaches an important lesson.

Theorem 16: $\mathbf{J}_{FS} \not\geq \mathbf{J}_{PS}$

Proof: Proof is by contradiction via counter-example. Let $W = 2$ with $\theta_1 = \theta_2 = 1/2$, and let $p_p = p_f = p$. Evaluate $\mathbf{1}_2^T (\mathbf{J}_{FS} - \mathbf{J}_{PS}) \mathbf{1}_2$, (the sum of each element of the matrix difference), which we expect to be nonnegative by assumption. It can be shown that this reduces to

$$2 - \frac{2q(1+q^2)}{1+q} - 2\frac{1+3q-4q^3}{1+2q} - q^2$$

which can be negative, e.g. when $q = 1/2$, a contradiction. ■

Using Lemma 27(ii) one can show that $\mathbf{J}_{FS} \not\geq \mathbf{J}_{PS}$ implies $\mathbf{J}_{PS}^{-1} \not\geq \mathbf{J}_{FS}^{-1}$ which in turn implies $\mathcal{I}_{FS}^+ \not\leq \mathcal{I}_{PS}^+$ (see Lemma 26). Since we earlier showed that PS is enormously worse than FS for variances, this result is surprising, particularly as experience shows that it is not difficult to find other examples for much larger W . We can explain this quandary as follows. When we focus on the variances in isolation we highlight the poor performance of PS. However, linear combinations such as $\sum_k a_k \theta_k$ bring in cross terms, which for PS are negative because of strong ambiguity in its observations. Strong correlations are a bad feature of a covariance matrix of an estimator, however if they are negative, they can in some cases cancel other positive terms resulting in a lower total variance. Hence, paradoxically, it is the poor behaviour of PS which prevents $\mathbf{J}_{FS} \geq \mathbf{J}_{PS}$ from holding.

The last example reveals that a ranking of methods based on a positive semidefinite comparison, although very desirable when true, is not possible in general. We therefore turn to a more generally applicable approach in the next section which we use for the remainder of the paper.

C. Variance Comparisons: a Partial Ranking

In this section we focus on comparing methods via the diagonal elements of \mathcal{I}^+ , corresponding to the optimal variances of θ_i estimates, just as in the figures above. As before, it is sufficient to consider the unconstrained covariance matrix \mathbf{J}^{-1} since $\mathcal{I}^+ = \mathbf{J}^{-1} - \boldsymbol{\theta}\boldsymbol{\theta}^T$ by (15). To the extent possible, we seek to establish a hierarchy between methods based on diagonal comparisons. Thanks to Theorem 15, which applies in particular to diagonal elements, we already have the answer in some cases.

We begin by comparing PS and PS+SYN. The examples given so far showed PS+SYN as enormously superior to PS. However, this is not always the case. A counterexample under the PPR normalization with $p = 0.05$ is given by

$$\boldsymbol{\theta} = \{\theta_1, \theta_1, \theta_{3 \leq k \leq 10} = \theta_1/100\}, \quad \theta_1 = 0.4808, \quad (30)$$

for which $D \approx 1.69$. As seen in Figure 7, PS outperforms PS+SYN for θ_9 and θ_{10} , but not otherwise. This makes sense given the bias of PS toward sampling large flows.

For these same parameters PS+SYN outperforms PS for all θ_k under ESR. This is typically the case. For example, with the above parameter set, PS+SYN outperforms PS on all sampling rates. While it may be true that PS outperforms PS+SYN under ESR in some rare situations, we believe this is not the case, and conjecture that PS+SYN defeats PS under ESR for all $\boldsymbol{\theta}$. For small p , we can prove that this is the case.

Theorem 17: Under PPR and ESR normalization, for small enough p , $(\mathbf{J}_{PS+SYN}^{-1})_{jj} \leq (\mathbf{J}_{PS}^{-1})_{jj}$ for any $\boldsymbol{\theta}$, $j = 1, \dots, W$.

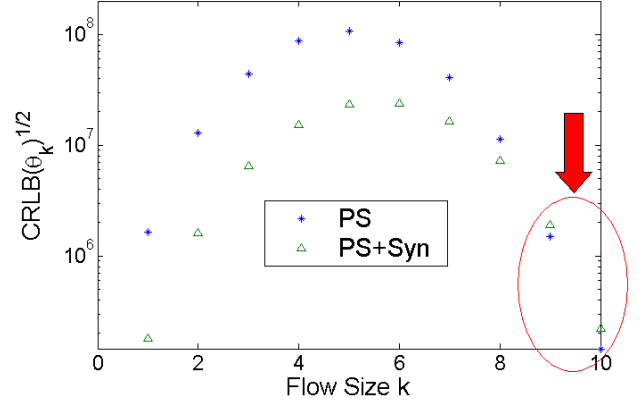


Fig. 7. PPR comparison between PS and PS+SYN with $p_p = 0.05$ for a particular distribution highly skewed to small flows.

We now consider the above comparison after both methods have benefitted from the use of sequence numbers.

Theorem 18: Under PPR and ESR normalization, for every $3 \leq j \leq W$, $(\mathbf{J}_{PS+SYN+SEQ}^{-1})_{jj} \leq (\mathbf{J}_{PS+SEQ}^{-1})_{jj}$.

For each of $j = 1$ and 2 counterexamples can be found for certain $\boldsymbol{\theta}$ and p , under both PPR and ESR. For example PS+SEQ has lower variance for θ_1 for the parameter set (30) with $p = 0.05$ under both PPR and ESR, and for θ_2 the family $\boldsymbol{\theta} = \{a, 1 - 2a, a\}$ with $a < 0.1$ gives examples for wide ranges of p , including as $p \rightarrow 0$ for a small enough. The counterexamples occur in atypical situations where θ_1 or θ_2 are large relative to other θ_j , and can be excluded if these are appropriately controlled. For example if $p < 1/2$ it can be shown that when $\theta_2/\theta_3 < q_p^2$, then PS+SEQ is worse under PPR (and hence ESR). Given that PS+SYN+SEQ defeats PS+SEQ except for perhaps θ_1 or θ_2 under atypical conditions, in general PS+SYN+SEQ can be regarded as superior.

The picture emerging from the above comparisons is that PS+SYN+SEQ is clearly superior to PS and PS+SEQ. Since PS+SYN+SEQ is just a special case of DS, we now consider this family in more detail. We focus on the ESR normalization which, apart from being far more important than PPR, is the key to our optimality result in Section VI. The following is one of our main results, a detailed characterization of the performance of DS under ESR. It can be shown that an analogous result does not hold for PPR.

Theorem 19: The diagonal elements $2 \leq j \leq W$ of \mathbf{J}_{DS}^{-1} under the ESR normalization are monotonically decreasing in

p_p . The property holds for $j = 1$ iff the condition

$$\theta_2 \geq \frac{D-1}{D} \theta_1 (1 - \theta_1) \quad (31)$$

is satisfied. Also, monotonicity holds when $D = 1$ or $W = 2$.

The above result shows, provided (31) is satisfied, that the variance bounds for each θ_k under DS drop as p_p increases. It follows that the optimal point p'_p in the DS family lies at $p'_p = 1$, that is flow sampling! The superiority of FS clearly demonstrates that the problem of inverting from imperfect sampled flows is so difficult that in the tradeoff between flow quality and quantity, quality wins convincingly.

The exception is for θ_1 when (31) fails, in which case the optimal DS lies to the left of $p_p = 1$. However, this only occurs when θ_2/θ_1 is extremely small, and even then in most cases $p'_p \approx 1$ unless D is also very large. Consider the region $p_p \rightarrow 1$. By solving for the optimal p_p^* using (41), we obtain (provided $D > 1$).

$$p_p^* = \sqrt{\frac{\theta_2}{(D-1)[\theta_1(1-\theta_1) - \theta_2]}}. \quad (32)$$

For p_p^* to be much smaller than 1, we require D to be large and the ratio θ_2/θ_1 very small (effectively, θ_2 being negligible). Such a scenario is very unlikely to occur in networks and in any case only affects θ_1 , and so it is reasonable to assume that the optimal DS corresponds to FS in practice.

Our next result compares the final method, SH, against FS.

Theorem 20: Under PPR and ESR normalization, for every $2 \leq j \leq W$, $(\mathbf{J}_{\text{FS}}^{-1})_{jj} \leq (\mathbf{J}_{\text{SH}}^{-1})_{jj}$. For $j = 1$ the condition holds for any θ for p sufficiently small. Sufficient conditions independent of p are either $W = 2$, or for $W > 2$,

$$\theta_2 \geq \theta_1(1 - \theta_1). \quad (33)$$

The proof for (33) involves bounds which are tight as $p \rightarrow 1$. Hence, when the condition is violated and p is large enough, SH can indeed outperform FS for θ_1 . An example is given in Figure 8 with $W = 6$ and $p_p = 0.4351$, which is large and reasonably close to $p = 0.5$. However, this effect is difficult to see in distributions with larger W as SH shifts most of its packet budget to larger flows, resulting in p_p being orders of magnitude smaller than p and the bounds leading to the sufficient condition becoming very loose. In all cases of interest therefore, FS can be regarded as superior to SH at all flow sizes.

Finally, we clarify the relationship between DS and SH more generally. DS is a two parameter family whose performance varies in a wide range. Indeed, Figure 5 already shows that it can perform better than SH at the ‘FS’ end of its range, and worse at the other end. The important observation is that in all cases where FS outperforms SH for a given θ , it follows from continuity of the CRLB in $(p_p, p_f(p))$ that there are members of the DS family other than FS which do likewise. Theorem 20 shows that this is true in almost all cases under ESR, allowing us to conclude that in general DS outperforms SH at all flow sizes.

The natural counterexample to this general rule is when (31) is satisfied so that the best DS can do is given by FS, and yet θ is such that SH defeats FS (for θ_1), implying that SH defeats all members of the DS family for θ_1 . This can occur if (33) is not satisfied and p is large. Furthermore there are cases where both SH and the optimal DS defeat FS, in which case the comparison is more complex. Since however this battle is contained within a very small and unimportant region of $(\theta, p, p_f(p))$ space, we do not characterise these counterexamples fully but instead conclude with the following result, which for the first time exhibits specific DS members (other than those in a neighbourhood of FS) which defeat SH.

Theorem 21: Under ESR normalization with rate p , if $0 \leq p_{p,\text{SH}} \leq \frac{pD-1}{D-1}$, then a sufficient condition for $(\mathbf{J}_{\text{DS}}^{-1})_{jj} \leq (\mathbf{J}_{\text{SH}}^{-1})_{jj}$ for every $1 \leq j \leq W$ is that $p_{p,\text{DS}}$ satisfies

$$p_{p,\text{SH}} \leq p_{p,\text{DS}}.$$

Summary: We have confirmed that a ranking of methods based on a positive semidefinite comparison is impossible, since even a diagonal comparison does not yield a simple hierarchy. Nonetheless, ignoring the counterexamples which sometimes occur for the CRLB variance for θ_1 and perhaps θ_2 , a clear overall picture emerges from the comparisons above. First, from Theorem 15, the utilization of sequence numbers yields consistent information theoretic dividends. In particular, PS need not be considered further since it has extremely poor performance and by Theorem 15 PS+SEQ is better. Theorems 17 and 18 show that the use of SYN sampling as a technique to eliminate bias is very powerful, which makes PS+SYN+SEQ the leading candidate, and focussed attention on its generalization, DS. Theorem 19 then shows that DS outperforms PS+SYN+SEQ under the ESR normalization (the most relevant one for estimation performance) provided $p_p > p_p^*$, and furthermore that FS is the favored member of DS. Theorem 20 shows that FS is also superior to SH. In conclusion, FS is the best sampling method for all flow sizes.

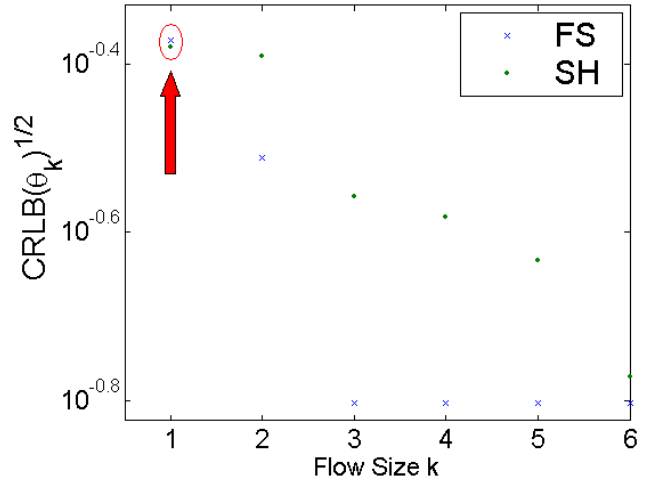


Fig. 8. PPR and ESR comparison between FS and SH with $p = 0.5$ for a particular distribution highly skewed to small flows.

This strongly motivates the adoption of sampling methods which approach FS as closely as possible, but which are efficiently implementable. This refocuses attention back to DS as it enjoys both of these properties.

V. ANALYSIS OF THE SEQ DIVIDEND

The utilization of sequence numbers provides additional ‘virtual’ packets for free. The extent to which this improves our information on θ clearly depends on parameters. For example, flows with only $k = 1$ or 2 packets are not helped at all. Intuitively (for example, under PS) it is the flows for which $k \gg 1/p_p$ which will receive the most benefit: since only the packets before the first and after the last physically sampled packets will not be recovered using sequence numbers, and on average there are $2/p_p$ of these since $1/p_p$ is the average gap between two physically sampled packets within a flow.

To quantify the benefit of sequence numbers better we define the *effective packet gain*, which is the ratio $r = \mathbb{E}[\tilde{N}_k]/\mathbb{E}[N_k]$, where \tilde{N}_k and N_k are the number of SEQ assisted and unassisted (physical) packets sampled respectively from a flow of size k . By definition, $R = \tilde{N}_k/N_k \geq 1$ and so $r \geq 1$, and asymptotically the gain saturates at $r = 1/p_p$.

For DS it can be shown (see [15] for more details) that to obtain a ratio α of this maximum gain, a flow must have a size of the order of $\frac{q_p(1+\alpha)}{p_p(1-\alpha)}$, suggesting that flows must have a size of roughly $1/p_p$ or more before the sequence number ‘information dividend’ effectively switches on. Although intuitively appealing, this is however quite misleading. There is a high variance associated with the random variable \tilde{N}_k , which for large flows under PS+SYN+SEQ goes as

$$\text{Var}(\tilde{N}_k) \approx \frac{4}{p_p} + k(k-2)p_p + (2k-3).$$

Therefore the gain R is not sharply concentrated around its mean and is highly dependent on the sampling parameter.

More generally, the sequence number dividend ‘switch’ is primarily about whether at least two packets are sampled. When p is small this is a very rare event, but can nonetheless be responsible for most of the available information about flow size. For example in Figure 3 the probability of sampling two packets is of the order of 10^{-6} , and yet the ratio of variance of PS to PS+SEQ is around 400 – even rare dividends are worth having! Thus even for smaller to medium sized flows (‘mice’ and ‘rabbits’), the use of sequence numbers provides a huge reduction in estimator variance, as shown by numerical and empirical evidence throughout this paper.

It is interesting to compare the SEQ dividend with SH, which implements a kind of extreme SEQ effect in the sense that if only a single packet is sampled, then all subsequent ones will be. This is a more powerful ‘switch’ than the SEQ mechanism which requires at least two packets. For large flows however, packets inferred under DS approach this level of packet recovery since, because the SYN packet is sampled, only a geometric number of packets will be missed at the end of the flow, compared to a geometric number at the beginning under SH. Of course, SH achieves this solely from its real packet budget whereas DS does not, which explains why DS

outperforms SH even in the latter’s area of strength, namely very large flows.

VI. COMPUTATIONAL ISSUES

A. Optimizing DS with Computational Constraints

From our result on the monotonic behaviour of ESR normalized DS, what is optimal in terms of information is clear: simply move down the ESR constraint curve to FS. However, there are other constraints from the resource side which will constrain the parts of the curve that will be accessible. The solution to the joint information/resource optimization problem is therefore clear: move as far down the ESR curve as possible. Our task now is to determine those constraints and hence the region in the (p_p, p_f) plane which is feasible.

The bottlenecks in the implementation of any sampling method are the memory access time and memory size. CPU processing is included in the memory access, since the CPU has to read out values from memory during lookup, performing modifications and write-backs when necessary. These tasks constitute the main portion of what is required of the CPU for measurement, as the measurement process is basically all about efficient counting [5]. This can be done for example with a hybrid SRAM-DRAM counter architecture [16], [17], [18], where a small amount of (fast but expensive) SRAM is used for the counter and its value periodically exported to a (cheaper but slower) DRAM store. With about 10 ns access time for SRAM, such a counter can be implemented even for OC-768 links which run at 40 Gbps.

We now consider how to optimize DS based on these constraints. Regarding flows, let T_{max} be the maximum flow table size measured in terms of flow records and λ_F be the flow arrival rate. As for packets, let C be the capacity of the link, P the size of the smallest packet (≈ 40 bytes), and τ the access time of memory (nominally DRAM).

Our simple analysis of the above bottleneck constraints is based on the following. In terms of packet arrivals we assume the worst case, namely the smallest size packets arriving back-to-back at line rate. By bounding the processing in such a case we guarantee that the front line of packet processing (which occurs at the highest speeds) is not under-dimensioned. In terms of flow arrivals we assume the ‘average case’ based on the average number of active flows. This can easily be made more conservative by replacing the average by some quantile to take into account fluctuations in flow arrivals. For simplicity, we ignore data export constraints from the measurement center to a central data collection center. We also do not consider rate adaptation based on the traffic condition although such schemes are compatible with DS.

Consider the processing of a single packet. The SYN bit is first tested to see which sampling parameter will apply, the cost of this is negligible, and if a packet is not sampled no further action is needed. Now consider the cost of a packet which is sampled. Each SYN packet which is sampled is inserted into the flow table. No prior lookup is necessary since it must be the first packet of its flow. Each non-SYN packet which is sampled must first perform a lookup of its flow-ID in the flow table to see if that flow is being tracked. If not it is discarded.

The cost of this wasteful per packet implementation of the ‘flow discarding’ step (inherent to any SYN based method such as DS) is not the bottleneck because the following case is the most expensive of all and is the one we model: a non-SYN packet which is sampled and whose flow *is* being tracked requires both a lookup followed by an update.

Using the above, the constraints are

$$p_f \leq \hat{p}_f = \min\left(\frac{T_{max}}{D\lambda_F}, \frac{P}{\tau C}\right), \quad p_p \leq \hat{p}_p = \frac{P}{2\tau C}. \quad (34)$$

The constraint $T_{max}/(D\lambda_F)$ ensures that the average number $D\lambda_F$ of active flows does not exceed the flow table size. Constraint $P/(\tau C)$ provides the worst case bound for per packet processing, for a single operation (insert or update). The factor of 2 that appears in the denominator for \hat{p}_p is to account for the worst case, in which a lookup *and* update is needed. The analysis is based on the use of a single per flow counter to track flow size. In practice, there may be more counters needed to track other quantities of interest, necessitating tighter constraints.

In the sequel, our examples consider traffic mixes that have manageable numbers of SYN packets, so that from (34), $\hat{p}_f = T_{max}/(D\lambda_F)$. This is done mainly to illustrate the relationship between the CRLB and parameter p_f . Indeed, at lower link speeds (OC-48 for example), p_f is determined by the number of flows arriving at the measurement point, rather than packet processing time. Secondly, to increase accuracy, from the previous discussions, we would like $p_f < p_p$, i.e. fewer, but better quality sampled flows.

The constraints form a simple region on the (p_p, p_f) plane that is convex, since it is rectangular with a corner at the origin. We want to minimize the variance for each θ_k subject to these constraints. Since the ESR curve is convex with respect to p_f and p_p , the optimal value must lie on the vertex of the convex constraint set [19, Corollary 32.3.1, p. 344]. For this to hold, we require that the optimum for θ_k on the ESR curve (Section IV-C) is outside the constrained region. This will be the case for all k for any reasonable traffic mix. Under such conditions, the solution is therefore $p_f = \hat{p}_f$ and $p_p = \hat{p}_p$.

We then have a relationship between the flow table size and link capacity and the diagonal elements of \mathbf{J}_{DS}^{-1} . Since it is apparent from (26) that the diagonals are dominated by $1/p_f$, by substituting the optimal solution, we have for $1 \leq j \leq W$,

$$(\mathbf{J}_{DS}^{-1})_{jj} = O\left(\frac{1}{\hat{p}_f}\right) = O\left(\frac{1}{T_{max}}\right).$$

Thus, with a larger table size (i.e. more memory available), variance of the estimator can be reduced.

As for the relation to capacity, we assume that C is large and use the approximation $\hat{q}_p \approx 1$. This can be justified considering that sampling is required beyond OC-3 link speeds.

The diagonals become

$$\begin{aligned} (\mathbf{J}_{DS}^{-1})_{11} &= \frac{\theta_1(1-\theta_1)}{\hat{p}_f} + \frac{1}{\hat{p}_f \hat{p}_p} \sum_{k=2}^W \theta_k + \theta_1^2 \\ (\mathbf{J}_{DS}^{-1})_{jj} &= \frac{\theta_j(1-\hat{p}_p \theta_j)}{\hat{p}_f \hat{p}_p} + \frac{1}{\hat{p}_f \hat{p}_p} \sum_{k=j+1}^W 2\theta_k + \theta_j^2 \\ (\mathbf{J}_{DS}^{-1})_{WW} &= \frac{\theta_W}{\hat{p}_f \hat{p}_p} - \frac{1}{\hat{p}_f} \theta_W^2 + \theta_W^2 \end{aligned}$$

which are approximately inversely proportional to \hat{p}_p , and hence are $O(C)$. We conclude that the variance of DS is inversely proportional to memory usage and proportional to the link capacity.

As an example, consider an OC-192 link with $D\lambda_F = 1 \times 10^6$ flows/sec, $T_{max} = 100,000$ and an access time of 100 ns for DRAM. Let us assume a further 100 ns is required for further processes (e.g. sequence number information). Thus, $\hat{p}_f = 0.1$ and $\hat{p}_p = 0.08$. With our numerical evaluation on the Leipzig-II trace (discussed in the following section), this would imply that the trace contain at least 9.5×10^8 original flows in to achieve a standard deviation of 10^{-8} or better. If we compare this to PS with a sampling rate of $p_p = 0.1$, we require a staggering 5.6×10^{44} flows to achieve the same performance!

To observe the dependence on memory and link capacity, now consider an OC-768 link instead with $T_{max} = 10,000$. This time, we have $\hat{p}_f = 0.01$ and $\hat{p}_p = 0.02$. The number of flows required to achieve the same standard deviation now increases to 8.8×10^9 which is still orders of magnitude better than PS.

If we consider the SRAM-DRAM architecture discussed earlier, with state-of-the-art SRAM having access times of about 5-10 ns, this can only increase the value of \hat{p}_p . Ideally, we still keep p_f low to reduce the number of flow entries while increasing p_p to approach FS performance. Continuing on from the previous example of the OC-192 link and assuming $\tau = 5$ ns, \hat{p}_p can now approach 1. For the OC-768 example, $\hat{p}_f = 0.01$ while $\hat{p}_p = 0.8$, giving huge performance gains.

B. Other Issues

Checking for the SYN bit can be done in a simple way by testing the payload type in the IP header and then verifying the presence of the SYN bit. This takes much less effort than deep packet inspection systems. Furthermore, sampling decisions can be implemented using precalculated values, much faster than a straight random number generator implementation.

We address the problem of flow table overload by using the method proposed by Estan et al. [20] by defining discrete measurement time bins, where sampled flows are exported at the end of each time bin. Consequently, overload of the flow table can be avoided at the cost of increased export rate.

VII. CASE STUDY ON INTERNET DATA

In this section we test the performance of DS on two traffic traces under the ESR normalization, and compare it to FS and its closest competitor, SH. We also further examine the

benefit of sequence numbers both in a fully empirical setting, and an idealised one. Since we work with real data, we require an unbiased estimator for each method. We propose closed-form estimators that achieve the CRLB asymptotically. We test their statistical performance by examining how close they approach the empirical CRLB on the traces we used. In general, the empirical results match the theoretical ones from earlier sections well.

A. Data Traces

We used two publicly available network traces, Leipzig-II [21] and Abilene-III [22], which are each collections of anonymized packet headers passing through a single router. Since the raw traffic is at packet level, specialized software such as CoralReef [23] are required to reconstruct flows. We modified the software so that only TCP flows were analyzed. Summary statistics of these traces appear in Table III. Both traces are unidirectional, presenting problems when constructing a sequence number function, as elaborated later.

There are many flows whose SYN packet is missing as they began before the measurement interval. To be consistent with our model assumptions, we remove these. A similar situation was encountered in [1]. Fortunately, such flows are in the minority, for example with Leipzig-II they account for only 18%. Furthermore, when sampling the trace, we assume an infinite timeout, that is, flows are expired at the end of the measurement interval. This is in accordance with the fact that we do not consider flow splitting. In practice, timeouts would split a flow, resulting underestimation of flow size.

Note that some flows may be malformed and have one or more SYN packets within the flow. Potentially, DS may sample one of these packets and treat it as a new flow, instead of part of a longer one. However, such cases are rare. In Leipzig-II, there are only 468 of such flows, or $\approx 0.02\%$, and there are none in Abilene-III. These flows were left in the trace.

Trace	Link Capacity	Active TCP Flows	Duration (hh:mm:ss)	D
Leipzig-II	50 Mbps	2,277,052	02:46:01	19.76
Abilene-III	10 Gbps	23,806,285	00:59:49	16.12

TABLE III
SUMMARY OF THE TRACES USED

The value of D in Table III is the actual average flow size. In our experiments, we truncate Leipzig-II to $W = 1000$ and Abilene-III to $W = 2000$, resulting in D being 1.94 and 7.65 packets for each trace respectively. Truncation is performed by discarding all flows with size above W , which ensures that the assumption $\theta_k > 0$, $k = 1, 2, \dots, W$ is met.

B. Closed-Form Unbiased Estimators

The analysis of earlier sections were centered on the CRLB, which bounds what is achievable by any unbiased estimator, but it is neither a construction of such an estimator nor even a proof of its existence. When working with real data an actual estimator is required, ideally one that achieves the previously computed CRLB. The maximum likelihood

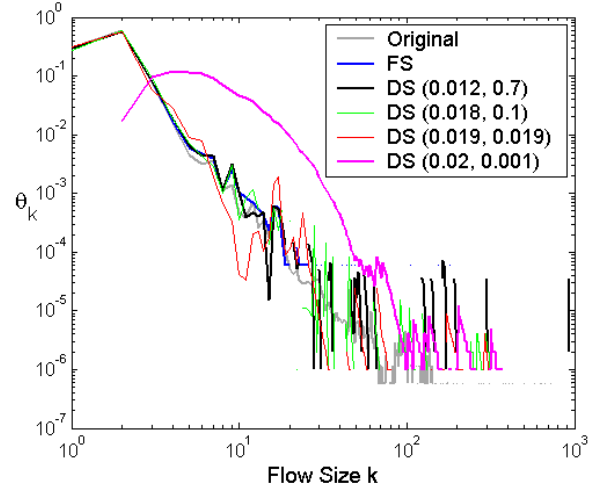


Fig. 9. Comparison of DS on Leipzig-II, using $W = 1000$ and varying parameters under ESR normalization with $p = 0.01$.

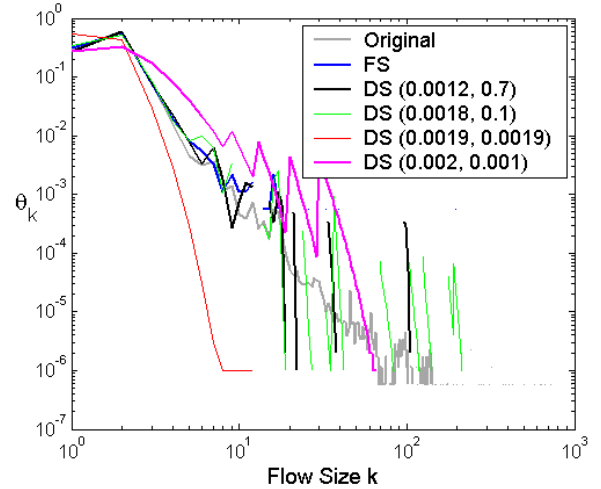


Fig. 10. Comparison of DS on Leipzig-II, using $W = 1000$ and varying parameters under ESR normalization with $p = 0.001$.

estimator (MLE) is an attractive candidate as it is *asymptotically efficient*, guaranteeing that its performance approaches the CRLB asymptotically [24]. However, the MLE, although unbiased asymptotically, is in general biased, especially in the small sample regime.

We propose the following estimator for FS and DS (or other SYN based methods such as PS+SYN):

$$\hat{\theta} = \frac{\tilde{\mathbf{B}}^{-1}}{N_f} [M'_1, M'_2, \dots, M'_W]^T. \quad (35)$$

This estimator has a natural interpretation as an empirical histogram based on observed sampled flow counts, inverted by $\tilde{\mathbf{B}}^{-1}$ to the original flow distribution. It is easy to see that it is unbiased since $\mathbb{E}[\hat{\theta}] = \tilde{\mathbf{B}}^{-1} \tilde{\mathbf{B}} \theta = \theta$. The matrix $\tilde{\mathbf{B}}^{-1}$ can be computed explicitly for these methods, as shown in Section III.

Our estimator is further motivated by its relation to the MLE, as we now show (all proofs are deferred to Appendix E).

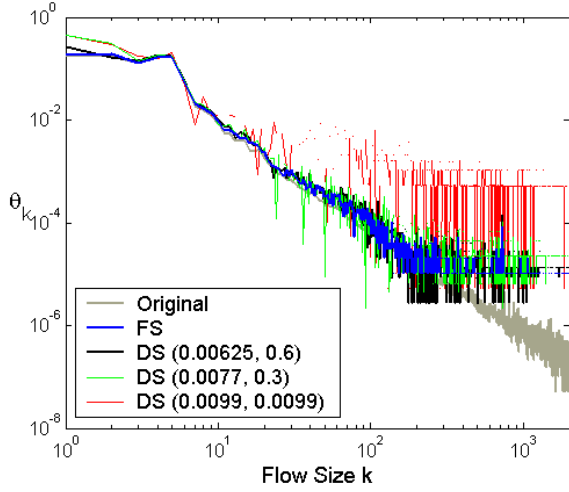


Fig. 11. Comparison of DS on Abilene-III, using $W = 2000$ and varying parameters under ESR normalization with $p = 0.005$.

Theorem 22: If the sampling matrix of a method satisfies $\mathbf{b}_0 = q\mathbf{1}_W$, then the MLE is

$$\hat{\theta} = \tilde{\mathbf{B}}^{-1} \left(\frac{p}{\sum_{j=1}^W M'_j} [M'_1, M'_2, \dots, M'_W]^T \right). \quad (36)$$

By rewriting $\sum_{j=1}^W M'_j$ as $N_f(1 - M'_0/N_f)$ and observing that M'_0/N_f converges to q with high probability (this follows from Hoeffding's inequality [25, p. 303]), the MLE (36) tends to our estimator (35), which therefore approaches the CRLB asymptotically. Our proposed estimator only obeys the constraint $\mathbf{1}_W^T \hat{\theta} = 1$ on average, while the estimator in Theorem 22 always obeys the constraint. (Note that (36) remains a viable estimator in the conditional framework, see Remark 32 in Appendix E.)

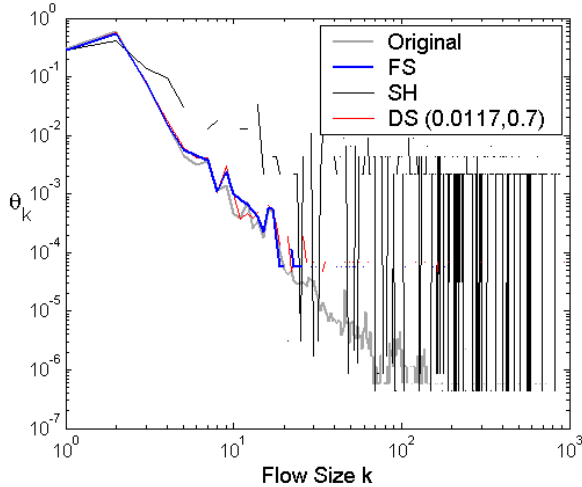


Fig. 12. Comparison of FS, SH ($p_r = 0.0002$) and DS ($p_f = 0.0117$, $p_p = 0.7$) on Leipzig-II, using $W = 1000$ and varying parameters under ESR normalization with $p = 0.01$.

The following applies to SH.

Theorem 23: The MLE for SH is given by

$$\hat{\theta}_{\text{SH}} = \frac{\tilde{\mathbf{B}}_{\text{SH}}^{-1}}{N_f} \cdot [p(M'_0 + M'_1), M'_2, \dots, M'_W]^T. \quad (37)$$

This time the MLE is unbiased (see Appendix E), so we use it directly on the data. It closely resembles the estimator (35), with a slight difference: for the estimate of θ_1 , the number of missing flows M'_0 plays a significant role. This can be interpreted as follows. Since SH works by geometrically skipping the first few packets in a flow, flows of size 1 are those most likely to have been entirely missed. Hence, the simplest way to incorporate a knowledge of M'_0 is to assume that it arises solely from evaporated flows originally of size 1.

The advantage of simple closed form estimators such as these, which only require a matrix multiplication, is that they eliminate the need for iterative estimation algorithms such as Expectation-Maximization (EM), often employed in the literature [1], [4], [8]. From a computational viewpoint, this is highly advantageous.

C. Testing with a Perfect Sequence Number Function

We begin our case study by testing DS with a *perfect* sequence number function, which returns the exact number of packets between two sampled packets. As we have access to the original unsampled flows, this is easily evaluated. Apart from being a benchmark for sequence number functions that may be designed in the future, a perfect function allows clean comparisons between alternative methods employing sequence numbers to be made.

In Figure 9, for an ESR of $p = 0.01$, values of p_p from $p_p = 1$ (equivalent to FS) steadily decreasing to $p_p = 0.001$ are shown (DS with $p_f = p_p = 0.019$ corresponds to PS+SYN+SEQ). As $p_p \rightarrow 1$ the performance vastly improves. Similar results were observed in Figure 10, where $p = 0.001$.

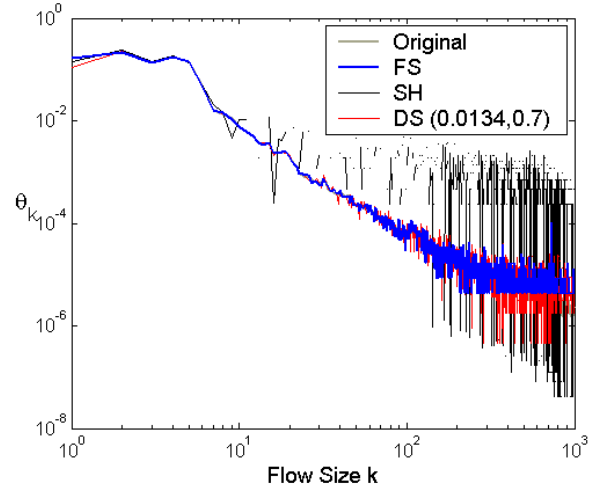


Fig. 13. Comparison of FS, SH ($p_r = 0.00018$) and DS ($p_f = 0.0134$, $p_p = 0.7$) on Abilene-III, using $W = 1000$ and varying parameters under ESR normalization with $p = 0.01$.

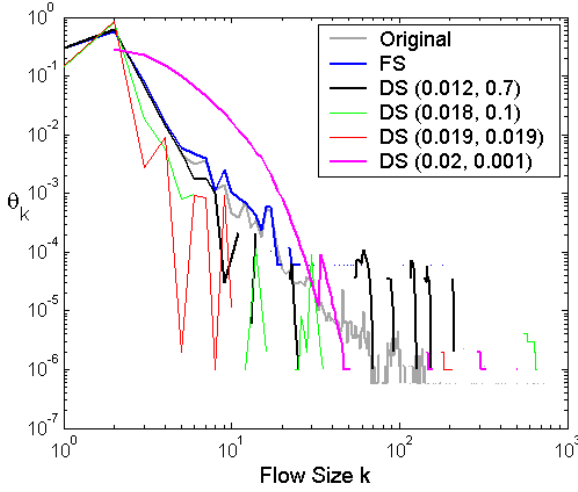


Fig. 14. Comparison of DS on Leipzig-II, using $W = 1000$ and varying parameters under ESR normalization with $p = 0.01$. An imperfect sequence number function is used here.

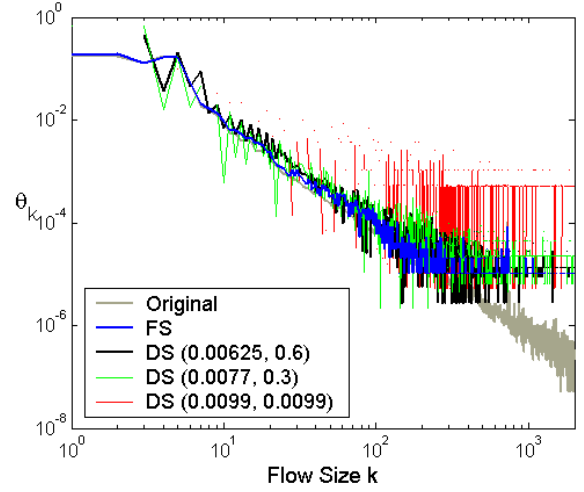


Fig. 16. Comparison of DS on Abilene-III, using $W = 2000$ and varying parameters under ESR normalization with $p = 0.005$. An imperfect sequence number function is used here.

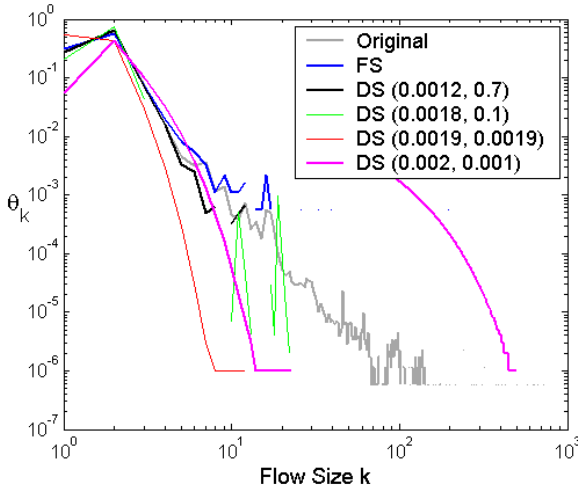


Fig. 15. Comparison of DS on Leipzig-II, using $W = 1000$ and varying parameters under ESR normalization with $p = 0.001$. An imperfect sequence number function is used here.

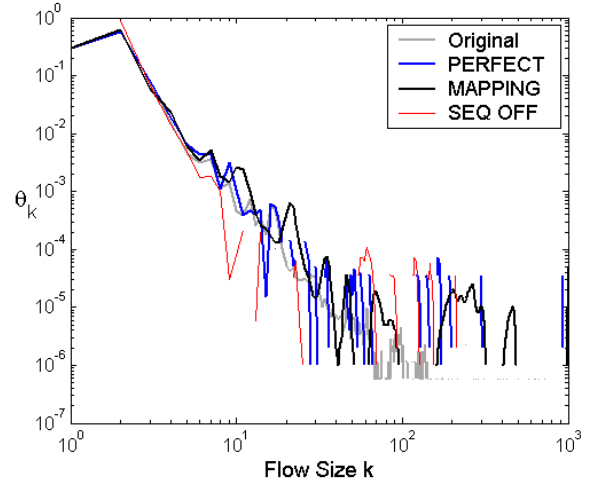


Fig. 17. Benefits of using sequence numbers. Three cases are shown: PERFECT uses a perfect sequence number function, MAPPING uses an imperfect function and SEQ OFF uses no sequence number information.

In all cases, a chronic lack of samples at the tail end of the distribution causes inaccurate estimates, with zero samples showing as discontinuities. FS holds to the true distribution the longest, as we expect, and the best DS performs similarly to it. With a much larger sample set in Figure 11, we can see better agreement between the estimates and the original distribution.

We also tested SH, as seen in Figures 12 and 13. To simplify the comparison across the traces, we truncate each to $W = 1000$, so that D becomes 1.94 and 6.45 packets for Leipzig-II and Abilene-III respectively, and use the same ESR value $p = 0.01$. In both cases the performance is much worse for SH than DS, which tracks FS and the true distribution well. In Leipzig-II the performance is very poor almost everywhere, while in Abilene-III the front end of the distribution, up to about flows of size 8, is estimated quite well, before deteriorating badly

at larger sizes. The good performance of SH at $j = 1$ can be attributed to the fact that this event is dominated by the sheer number of original flows of size one. Both FS and DS tracks the distribution much better since both methods have sampled flows of far superior quality to SH.

D. Testing with an Imperfect Sequence Number Function

We now test DS with an imperfect sequence number function. Our function is similar to that outlined in [4]. However, as we do not have statistics of popular TCP payload sizes available, we infer the most likely payload size as follows. If the sequence number difference is divisible by a popular payload size (for example, 1460 bytes), we take this as the most likely payload size. Otherwise, we use the average payload size. This function is subject to errors, especially when a flow has variable payloads, however it suffices for

our purposes.

In addition to TCP sequence numbers, we exploit IPID numbers. As mentioned in [4], the IPID field of Linux machines is incremented sequentially for each TCP flow every time a packet in the flow is transmitted. Given that the majority of web-servers on the Internet are Linux-based, we exploit IPID numbers to check the accuracy of our estimate when a flow has packets with variable payloads.

Furthermore, the unidirectional nature of the trace presents a significant challenge. As one side in a TCP connection usually transmits more data than the other, some sampled flows may consist mainly of TCP ACK packets, which means that they will have zero-byte TCP payloads. In this case sequence numbers may not be incremented at all, and so will not provide information about the number of bytes transmitted. A solution is to use the TCP acknowledgement numbers instead to infer the number of bytes transmitted from the opposite direction, which would yield an estimate of the number of packets in the TCP ACK flow. This may not be the most ideal solution, as this method would be susceptible to delayed ACKs, thus underestimating the size of the flow.

Modern web browsers rely on maintaining TCP connections rather than initiating new connections, which require more memory. This is the persistent HTTP protocol. The prevalence of this protocol amongst web browsers presents a challenge, since empty payload packets are periodically sent to keep the connection alive. These packets do not increment the sequence number. The best we can do in such cases is to infer using IPID numbers, or possibly counting ACK packets coming in from the other direction, if bidirectional information is available.

Even with the imperfect and relatively simple sequence number function used here, results are consistent with theory. Figures 14 and 15 illustrate this. In both cases, the imperfection of the function affects the accuracy of DS, but not to a large degree. A similar observation applies to the Abilene-III trace, shown in Figure 16, where the artifacts due to the imperfect function (the sawtooth pattern) are clearly seen.

Finally, Figure 17 illustrates the effect of using sequence numbers in recovering the flow size distribution. The three cases shown are for DS with parameters $p_f = 0.00117$ and $p_p = 0.7$, with an ESR of $p = 0.01$. The PERFECT case is when DS is given a perfect sequence number function, MAPPING when using our sequence number function, and SEQ OFF when no sequence numbers were used at all. It is apparent that using sequence numbers, even with an imperfect function, provides significantly more information to an estimator.

E. Empirical estimator variance

Here, we see how closely the estimator variance matches the CRLB by computing the *observed Fisher information* [26] of the estimator in Figure 18. To improve readability, smoothing was applied to the tail end of the observed Fisher information where samples are scarce using a simple moving average filter with a window size of 100. Even when using an imperfect sequence number function, the variance of the estimator closely matches the CRLB, effectively proving the benefit of sequence numbers.

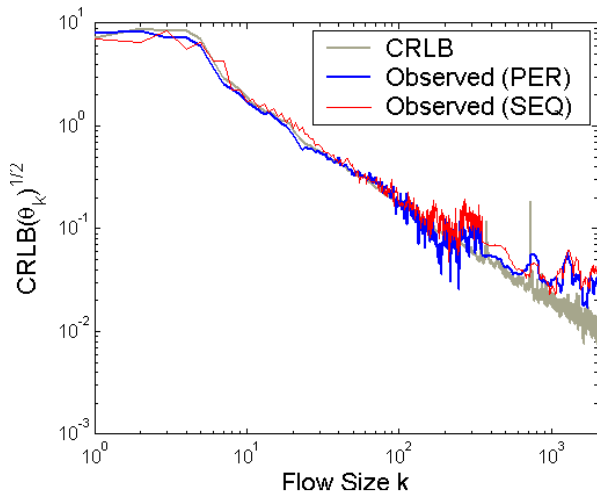


Fig. 18. CRLB versus observed MLE variance of DS with $p_f = 0.00625$, $p_p = 0.6$ on Abilene-III.

VIII. CONCLUSIONS AND FUTURE WORK

We have re-examined the question of sampling for flow size estimation in the context of TCP flows from a theoretical point of view. We used the Fisher information to examine the inherent potential of a number of sampling methods. Most of these had been examined previously, but we showed how the usual conditional framework can be made unconditional and thereby simplified, which actually changes the sampling methods themselves and their performance. The new framework led to a number of new rigorous results regarding the performance of sampling methods which we studied under two different normalizations. It also enabled flow sampling to be compared to methods using TCP sequence numbers and Sample and Hold for the first time, and we showed that it is far superior to them, except in very special cases which are not important for network measurement applications.

We introduced a new two parameter family of methods, Dual Sampling, which allows the statistical benefits of flow sampling to be traded off against the computational advantages of packet sampling. We discussed how, as an unbiased ‘Hold and Sample’ method, it differs from Sample and Hold, and proved that it is superior to it. We argue that the scheme is implementable and offers an efficient way of approaching flow sampling in practice to the extent possible. We also proposed closed-form unbiased estimators for SYN-based methods and SH which asymptotically achieve the CRLB, saving computational time in the estimation stage.

We performed a case study of Dual Sampling and Sample and Hold on two Internet data traces using our proposed estimators, and found results entirely consistent with the theoretical predictions, despite the fact that the function which maps sequence numbers to packet counts introduces a new source of error, and was not highly optimized. Although there is high variation at the tail end of the distribution, our proposed estimator closely matches the CRLB.

In future work, we intend to search over the space of all possible sampling matrices to find an optimal sampling method

for flow size estimation, and to compare it to flow sampling. Our framework is applicable to other traffic metrics such as anomaly detection and a future direction is to extend the work to those areas. It would also be of interest to improve the sequence number mapping function, and also to explore using our approach for the direct estimation of the byte size of flows, for which sequence numbers are more naturally suited, rather than the packet size. Finally, we will develop a more detailed case for DS and its implementability in high speed routers.

APPENDIX A OUR MATRIX LEMMAS

A. General

Lemma 24: The matrix $\tilde{\mathbf{J}}(\boldsymbol{\theta})$ and its inverse $\tilde{\mathbf{J}}(\boldsymbol{\theta})^{-1}$ are symmetric and positive definite.

Proof: For simplicity we omit the $\boldsymbol{\theta}$ dependencies. Recall that $\tilde{\mathbf{J}} = \tilde{\mathbf{B}}^T \tilde{\mathbf{D}} \tilde{\mathbf{B}}$ where $\tilde{\mathbf{D}}$ is a real diagonal positive definite matrix with $(\tilde{\mathbf{D}})_{jj} = d_j^{-1}$ and $\text{rank}(\tilde{\mathbf{D}}) = W$. Matrix $\tilde{\mathbf{B}}$ is a $W \times W$ matrix and $\text{rank}(\tilde{\mathbf{B}}) = W$. It follows that an inverse exists for both $\tilde{\mathbf{D}}$ and $\tilde{\mathbf{B}}$. Hence, an inverse also exists for $\tilde{\mathbf{J}}$ since $\tilde{\mathbf{J}}^{-1} = \tilde{\mathbf{B}}^{-1} \tilde{\mathbf{D}}^{-1} (\tilde{\mathbf{B}}^T)^{-1}$.

An equivalent expression for $\tilde{\mathbf{J}}$ is

$$\tilde{\mathbf{J}} = \tilde{\mathbf{B}}^T \tilde{\mathbf{D}}^{1/2} \tilde{\mathbf{D}}^{1/2} \tilde{\mathbf{B}} = (\tilde{\mathbf{D}}^{1/2} \tilde{\mathbf{B}})^T (\tilde{\mathbf{D}}^{1/2} \tilde{\mathbf{B}}) \quad (38)$$

since $\tilde{\mathbf{D}}^{1/2}$ is symmetric.

We can now show that $\tilde{\mathbf{J}}$ is symmetric. We have

$$\begin{aligned} \tilde{\mathbf{J}}^T &= [(\tilde{\mathbf{D}}^{1/2} \tilde{\mathbf{B}})^T (\tilde{\mathbf{D}}^{1/2} \tilde{\mathbf{B}})]^T \\ &= (\tilde{\mathbf{D}}^{1/2} \tilde{\mathbf{B}})^T (\tilde{\mathbf{D}}^{1/2} \tilde{\mathbf{B}}) = \tilde{\mathbf{J}}. \end{aligned}$$

The form $(\tilde{\mathbf{D}}^{1/2} \tilde{\mathbf{B}})^T (\tilde{\mathbf{D}}^{1/2} \tilde{\mathbf{B}})$ is at least positive semidefinite (Theorem 28). However, since $\tilde{\mathbf{J}}^{-1} = \tilde{\mathbf{B}}^{-1} \tilde{\mathbf{D}}^{-1} (\tilde{\mathbf{B}}^T)^{-1}$, $\tilde{\mathbf{J}}$ is invertible, it is also positive definite. By definition of symmetric, positive definite matrices, its inverse is also symmetric, positive definite. ■

Lemma 25: The unconstrained Fisher information matrix $\mathbf{J}(\boldsymbol{\theta})$ and its inverse $\mathbf{J}(\boldsymbol{\theta})^{-1}$ are symmetric and positive definite.

Proof: Recall that $\mathbf{J} = \mathbf{B}^T \mathbf{D} \mathbf{B}$ where \mathbf{D} is a real diagonal positive definite matrix with $(\mathbf{D})_{jj} = d_j^{-1}$ and $\text{rank}(\mathbf{D}) = W + 1$. Matrix \mathbf{B} is a $(W + 1) \times W$ matrix and $\text{rank}(\mathbf{B}) = W$. An equivalent expression for \mathbf{J} is

$$\mathbf{J} = \mathbf{B}^T \mathbf{D}^{1/2} \mathbf{D}^{1/2} \mathbf{B} = (\mathbf{D}^{1/2} \mathbf{B})^T (\mathbf{D}^{1/2} \mathbf{B})$$

since $\mathbf{D}^{1/2}$ is symmetric. Now

$$\begin{aligned} \mathbf{J}^T &= [(\mathbf{D}^{1/2} \mathbf{B})^T (\mathbf{D}^{1/2} \mathbf{B})]^T \\ &= (\mathbf{D}^{1/2} \mathbf{B})^T (\mathbf{D}^{1/2} \mathbf{B}) = \mathbf{J}. \end{aligned}$$

From Theorem 28, $(\mathbf{D}^{1/2} \mathbf{B})^T (\mathbf{D}^{1/2} \mathbf{B})$ is positive semidefinite. Moreover, \mathbf{J} is invertible by Proposition 2, implying it is positive definite. By definition of symmetric, positive definite matrices, its inverse is also symmetric and positive definite. ■

Lemma 26: $\mathbf{J}_1 \geq \mathbf{J}_2$ if and only if $\mathcal{I}_1^+ \leq \mathcal{I}_2^+$.

Proof: Since by Lemma 27(ii), $\mathbf{J}_1 \geq \mathbf{J}_2$ iff $\mathbf{J}_1^{-1} \leq \mathbf{J}_2^{-1}$, then by definition of positive semidefinite matrices, $\mathbf{x}^T \mathbf{J}_1^{-1} \mathbf{x} \leq \mathbf{x}^T \mathbf{J}_2^{-1} \mathbf{x}$. Hence, this implies $\mathbf{x}^T (\mathbf{J}_1^{-1} - \boldsymbol{\theta} \boldsymbol{\theta}^T) \mathbf{x} \leq \mathbf{x}^T (\mathbf{J}_2^{-1} - \boldsymbol{\theta} \boldsymbol{\theta}^T) \mathbf{x}$, implying $\mathbf{J}_1 \geq \mathbf{J}_2$. ■

B. Proof of Theorem 4

Let $\mathbf{E} = (1/d_0) \mathbf{b}_0 \mathbf{b}_0^T$ from (10). It has rank 1 and is therefore positive semidefinite since its eigenvalues are $\text{tr}(\mathbf{E}) = \sum_{k=1}^W b_{0k}^2/d_0$ with multiplicity 1 and 0 with multiplicity $W - 1$. It follows from Lemma 29 that $\mathbf{J}(\boldsymbol{\theta}) \geq \tilde{\mathbf{J}}(\boldsymbol{\theta})$ since $\mathbf{J}(\boldsymbol{\theta}) = \mathbf{E} + \tilde{\mathbf{J}}(\boldsymbol{\theta})$, and from Lemma 27, this implies that $\tilde{\mathbf{J}}^{-1}(\boldsymbol{\theta}) - \mathbf{J}^{-1}(\boldsymbol{\theta}) \geq \mathbf{0}_{W \times W}$ and therefore $\mathbf{J}^{-1}(\boldsymbol{\theta}) \leq \tilde{\mathbf{J}}^{-1}(\boldsymbol{\theta})$. Equality can only hold iff $\mathbf{E} = \mathbf{0}_{W \times W}$ since this is the only case where all eigenvalues of \mathbf{E} are zero. This implies that $\mathbf{b}_0 = \mathbf{0}_{W \times 1}$ is required for equality, that is that no flow can ‘evaporate’.

APPENDIX B OTHER MATRIX LEMMAS

We collect some useful results required in this paper here. This first result comes from [12].

Lemma 27: Let \mathbf{A} be a $n \times n$ symmetric positive definite matrix and \mathbf{B} an $n \times n$ positive definite matrix. Then

- (i) If $\mathbf{B} - \mathbf{A}$ is positive definite, then so is $\mathbf{A}^{-1} - \mathbf{B}^{-1}$,
- (ii) If $\mathbf{B} - \mathbf{A}$ is symmetric and positive semidefinite (implying \mathbf{B} is symmetric), then $\mathbf{A}^{-1} - \mathbf{B}^{-1} \geq 0$.

The following theorem appears in [27, Theorem 6.3, p. 161].

Lemma 28: The following statements are equivalent:

- (i) \mathbf{A} is positive semidefinite;
- (ii) $\mathbf{A} = \mathbf{B}^* \mathbf{B}$ for some matrix \mathbf{B} , where \mathbf{B}^* is the conjugate transpose of \mathbf{B} .

The following result gives more properties of positive semidefinite matrices [27, Theorem 6.5, p. 166].

Lemma 29: Let $\mathbf{A} \geq 0$ and $\mathbf{B} \geq 0$ have same size. Then

- (i) $\mathbf{A} + \mathbf{B} \geq 0$,
- (ii) $\mathbf{A}^{1/2} \mathbf{B} \mathbf{A}^{1/2} \geq 0$,
- (iii) $\text{tr}(\mathbf{A} \mathbf{B}) \leq \text{tr}(\mathbf{A}) \text{tr}(\mathbf{B})$,
- (iv) the eigenvalues of $\mathbf{A} \mathbf{B}$ are all nonnegative.

The matrix inversion lemma (also known as Woodbury’s formula) can be found in [12, Theorem 18.2.8, p. 424].

Lemma 30 (Matrix Inversion Lemma): Let \mathbf{R} be a $n \times n$ matrix, \mathbf{S} a $n \times m$ matrix, \mathbf{T} a $m \times m$ matrix, and \mathbf{U} a $m \times n$ matrix. Suppose that \mathbf{R} and \mathbf{T} are nonsingular. Then,

$$(\mathbf{R} + \mathbf{S} \mathbf{T} \mathbf{U})^{-1} = \mathbf{R}^{-1} - \mathbf{R}^{-1} \mathbf{S} (\mathbf{T}^{-1} + \mathbf{U} \mathbf{R}^{-1} \mathbf{S})^{-1} \mathbf{U} \mathbf{R}^{-1}.$$

The data processing inequality for Fisher information from [28] is as follows.

Theorem 31: If $\Theta \rightarrow Y \rightarrow X$ satisfies a relation of the form $f(y, x | \Theta) = f_\Theta(y) f(x | y)$ (i.e. the conditional distribution of X given Y is independent of Θ), then $\mathbf{J}_X(\Theta) \leq \mathbf{J}_Y(\Theta)$ with the deterministic version being $\mathbf{J}_{\gamma(Y)}(\Theta) \leq \mathbf{J}_Y(\Theta)$. Equality holds if $\gamma(Y)$ is a *sufficient statistic* relative to the family $f_\Theta(y)$, i.e. $\Theta \rightarrow \gamma(Y) \rightarrow Y$ forms a Markov chain.

APPENDIX C SAMPLING METHODS

A. Proof of equation (17)

Expanding $(1 + (-1))^k$ leads to the useful identity

$$\sum_{\ell=1}^k (-1)^{k-\ell} \binom{k}{\ell} = (-1)^{k-1}. \quad (39)$$

Using (12) and $b'_{jk} = (-1)^{k-j} \binom{k}{j} q_p^{k-j} p_p^{-k}$ for PS, we have

$$(\mathbf{J}^{-1})_{jj} = \sum_{k=j}^W \binom{k}{j}^2 q_p^{2(k-j)} p_p^{-2k} d_k - \frac{(\sum_{k=j}^W \sum_{\ell=1}^k (-1)^{2k-j-\ell} \binom{k}{j} \binom{k}{\ell} q_p^{2k-j} p_p^{-2k} d_k)^2}{d_0 + \sum_{k=1}^W d_k (q_p^k p_p^{-k} \sum_{\ell=1}^k (-1)^{k-\ell} \binom{k}{\ell})^2}.$$

The denominator can be simplified as follows:

$$\begin{aligned} d_0 + \sum_{k=1}^W d_k \left(q_p^k p_p^{-k} \sum_{\ell=1}^k (-1)^{k-\ell} \binom{k}{\ell} \right)^2 \\ = d_0 + \sum_{k=1}^W d_k q_p^{2k} p_p^{-2k} \left(\sum_{\ell=1}^k (-1)^{k-\ell} \binom{k}{\ell} \right)^2 \\ = d_0 + \sum_{k=1}^W (-1)^{2k-2} d_k q_p^{2k} p_p^{-2k} \\ = \sum_{k=0}^W q_p^{2k} p_p^{-2k} d_k, \end{aligned}$$

where we used identity (39) in the second line.

Similarly, the numerator can be simplified:

$$\begin{aligned} & \left(\sum_{k=j}^W \sum_{\ell=1}^k (-1)^{2k-j-\ell} \binom{k}{j} \binom{k}{\ell} q_p^{2k-j} p_p^{-2k} d_k \right)^2 \\ &= \left(\sum_{k=j}^W d_k \binom{k}{j} (-1)^{k-j} q_p^{2k-j} p_p^{-2k} \sum_{\ell=1}^k (-1)^{k-\ell} \binom{k}{\ell} \right)^2 \\ &= \left(\sum_{k=j}^W d_k \binom{k}{j} (-1)^{k-j} q_p^{2k-j} p_p^{-2k} (-1)^{k-1} \right)^2 \\ &= \left(\sum_{k=j}^W (-1)^{2k-j-1} d_k \binom{k}{j} q_p^{2k-j} p_p^{-2k} \right)^2, \end{aligned}$$

where (39) is used in the third line. This proves (17).

APPENDIX D COMPARISONS

A. Proof of Theorem 14

Let Z denote any method except SH, and Y the complete outcome (i.e. the vector of SEQ numbers and the SYN variable in the richest case) of sampling with Z using parameter p_1 (for DS $(p_{p,1}, p_{f,1})$). Now sample Y using Z with parameter p_2/p_1 to form X (for DS $(p_{p,2}/p_{p,1}, p_{f,2}/p_{f,1})$). Since X is a function only of Y and the new sampling, the DPI for Fisher applies (Theorem 31) (and $X \subseteq Y$). Furthermore, it is easy to see that X is statistically equivalent to the outcome of sampling the original data using Z with probability $p_1(p_2/p_1) = p_2$ (for DS $(p_{p,1}(p_{p,2}/p_{p,1}), p_{f,1}(p_{f,2}/p_{f,1})) = (p_{p,2}, p_{f,2})$). It follows that $\mathbf{J}_Z(p_1) \geq \mathbf{J}_Z(p_2)$. Equality holds iff $p_2 = p_1$ (for DS $p_{p,1} = p_{p,2}$ and $p_{f,1} = p_{f,2}$) implying $X = Y$, since sampling inherently discards information.

The above proof does not apply to SH as it does not have a closure property, meaning that X is not equivalent to

applying SH with some p_p . We turn instead to a direct method, exploiting the tridiagonal structure of $\mathbf{J}_{\text{SH}}^{-1}$.

Denote $(\mathbf{J}_{\text{SH}}^{-1})_{ij} = a_{i,j}$, and consider the quadratic form $\mathbf{x}^T \mathbf{J}_{\text{SH}}^{-1} \mathbf{x}$ for any non-zero vector $\mathbf{x} \in \mathbb{R}^W$:

$$\begin{aligned} \mathbf{x}^T \mathbf{J}_{\text{SH}}^{-1} \mathbf{x} &= (a_{1,1} + a_{1,2})x_1^2 + (a_{W,W} + a_{W-1,W})x_W^2 \\ &+ \sum_{j=2}^{W-1} (a_{j,j} + a_{j-1,j} + a_{j,j+1})x_j^2 \\ &+ \sum_{j=1}^W (-a_{j,j+1})(x_j - x_{j+1})^2. \end{aligned} \quad (40)$$

From Lemma 5 we know $\mathbf{J}^{-1} \mathbf{1}_W = \boldsymbol{\theta}$. It follows that

$$\begin{aligned} (a_{1,1} + a_{1,2}) &= \theta_1 \\ (a_{j,j} + a_{j-1,j} - a_{j,j+1}) &= \theta_j, \quad j = 2, \dots, W-1 \\ (a_{W,W} + a_{W-1,W}) &= \theta_W \end{aligned}$$

which are each independent of p_p . Consider then the term involving $-a_{j,j+1} = p_p^{-2} q_p d_{j+1} = \frac{1}{p_p} \sum_{k=j+1}^W q_p^{k-j} \theta_k$, $1 \leq j < W$. Differentiating with respect with p_p , we obtain

$$-\frac{1}{p_p^2} \sum_{k=j+1}^W q_p^{k-j} \theta_k - \frac{1}{p_p} \sum_{k=j+1}^W (k-j) q_p^{k-j-1} \theta_k.$$

Since this is negative, it follows that $\mathbf{x}^T \mathbf{J}_{\text{SH}}^{-1} \mathbf{x}$ decreases monotonically in p_p , and so $\mathbf{J}_{\text{SH}}^{-1}(p_1) \leq \mathbf{J}_{\text{SH}}^{-1}(p_2)$ for any $p_1 \geq p_2$. The result then follows by Lemma 27(ii).

B. Proof of Theorem 17

We first consider PPR normalization, where $p_p = p$ for both methods. Recall from (17) that

$$\begin{aligned} (\mathbf{J}_{\text{PS}}^{-1})_{jj} &= \sum_{k=j}^W \binom{k}{j}^2 q^{2(k-j)} p^{-2k} d_{k,\text{PS}} \\ &- \frac{(\sum_{k=j}^W (-1)^{2k-j-1} d_{k,\text{PS}} \binom{k}{j})^2 q^{2k-j} p^{-2k}}{\sum_{k=0}^W q^{2k} p^{-2k} d_{k,\text{PS}}} \\ &\geq \underbrace{\sum_{k=j}^W \binom{k}{j}^2 q^{2(k-j)} p^{-2k} d_{k,\text{PS}}}_{T_1} - q^{-2j} (W-j+1) W! \end{aligned}$$

where the lower bound follows by, in the second term, dropping the first j terms in the denominator and upper bounding $\binom{k}{j}$ by $W!$. Also, from (18),

$$(\mathbf{J}_{\text{PS}+\text{SYN}}^{-1})_{jj} = \underbrace{\sum_{k=j}^W \binom{k-1}{j-1}^2 q^{2(k-j)} p^{-2k} d_{k,\text{PS}+\text{SYN}}}_{T_2} - \frac{q}{p} \theta_j^2.$$

Since $\binom{k}{j} = \binom{k-1}{j-1} + \binom{k-1}{j} \geq \binom{k-1}{j-1}$,

$$d_{j,\text{PS}} \geq \sum_{k=j}^W \binom{k-1}{j-1} p^j q^{k-j} \theta_k = d_{j,\text{PS}+\text{SYN}},$$

implying that $T_1 \geq T_2$.

Now under the limit $p \rightarrow 0$, for $j \geq 1$, the second term for PS becomes $-(W - j + 1)W!$ which is finite, whereas for PS+SYN, $-\frac{q}{p}\theta_j^2 \approx -\frac{1}{p}\theta_j^2 \rightarrow -\infty$. It follows that $(\mathbf{J}_{\text{PS+SYN}}^{-1})_{jj} \leq (\mathbf{J}_{\text{PS}}^{-1})_{jj}$. Since $p_p > p$ for PS+SYN under ESR, the result also holds for ESR by Theorem 14.

C. Proof of Theorem 18

We begin by examining the simplest case, where $j = W$. We have

$$(\mathbf{J}_{\text{PS+SEQ}}^{-1})_{WW} = \frac{\theta_W}{p_p^2}$$

and

$$(\mathbf{J}_{\text{PS+SYN+SEQ}}^{-1})_{WW} = \frac{\theta_W}{p_p^2} - \frac{q_p}{p_p} \theta_W^2,$$

and so $(\mathbf{J}_{\text{PS+SYN+SEQ}}^{-1})_{WW} \leq (\mathbf{J}_{\text{PS+SEQ}}^{-1})_{WW}$ under PPR.

Now consider the case $3 \leq j \leq W - 1$. Let $d_{Q,j}$ and $d_{SQ,j}$ be the proportion of sampled flows of size j after sampling by PS+SEQ and PS+SYN+SEQ respectively. For $j \geq 1$, a straightforward comparison would yield $d_{Q,j} \geq d_{SQ,j}$. Intuitively, since more flows are discarded in the PS+SYN+SEQ scheme, the proportion of sampled flows must be less than the proportions of PS+SEQ. Therefore, expressing the diagonals in terms of $d_{Q,j}$ and $d_{SQ,j}$, we have for $3 \leq j \leq W - 1$,

$$(\mathbf{J}_{\text{PS+SEQ}}^{-1})_{jj} = p_p^{-4} d_{Q,j} + 4q_p^2 p_p^{-4} d_{Q,j+1} + q_p^4 p_p^{-4} d_{Q,j+2}$$

and

$$\begin{aligned} (\mathbf{J}_{\text{PS+SYN+SEQ}}^{-1})_{jj} &= p_p^{-4} d_{SQ,j} + q_p^2 p_p^{-4} d_{SQ,j+1} - q_p p_p^{-1} \theta_j^2 \\ &\leq p_p^{-4} d_{SQ,j} + q_p^2 p_p^{-4} d_{SQ,j+1}, \end{aligned}$$

which, by a direct comparison, shows $(\mathbf{J}_{\text{PS+SYN+SEQ}}^{-1})_{jj} \leq (\mathbf{J}_{\text{PS+SEQ}}^{-1})_{jj}$. Similarly, the ESR comparison follows since the sampling rate for PS+SYN+SEQ must increase, thereby reducing its CRLB.

D. Proof of Theorem 19

First consider the simplest case, $j = W$. By substituting (27) into (26) and then differentiating w.r.t p_p

$$\frac{d}{dp_p} (\mathbf{J}_{\text{DS}}^{-1})_{WW} = -\frac{\theta_W}{p_p^2 p D} - \frac{(D-1)\theta_W^2}{p D} < 0,$$

implying $(\mathbf{J}_{\text{DS}}^{-1})_{WW}$ is monotonically decreasing with p_p .

For $2 \leq j \leq W - 1$ the derivative $\frac{d}{dp_p} (\mathbf{J}_{\text{DS}}^{-1})_{jj}$ is given by

$$\begin{aligned} &-\frac{\theta_j}{p_p^2 p D} - \frac{(D-1)\theta_j^2}{p D} - \frac{1}{p_p^2 p D} \left(\sum_{k=j+1}^W q_p^{k-j} (1 + q_p) \theta_k \right) \\ &- \frac{1}{p_f p_p} \left(\sum_{k=j+1}^W q_p^{k-j-1} ((k-j) + (k-j+1)q_p) \theta_k \right), \end{aligned}$$

which is negative since each term is negative for $0 < p_p \leq 1$.

Finally, for $j = 1$ we have

$$\begin{aligned} \frac{d}{dp_p} (\mathbf{J}_{\text{DS}}^{-1})_{11} &= \frac{D-1}{p D} \theta_1 (1 - \theta_1) - \frac{1}{p_p^2 p D} \left(\sum_{k=2}^W q_p^{k-1} \theta_k \right) \\ &- \frac{p_p (D-1) + 1}{p_p p D} \left(\sum_{k=2}^W (k-1) q_p^{k-2} \theta_k \right). \quad (41) \end{aligned}$$

For small values of p_p the expression is dominated by terms in $1/p_p$ and is therefore again negative, but as the first term is positive, for large p_p it may change sign. It is not hard to show that $\frac{d}{dp_p} (\mathbf{J}_{\text{DS}}^{-1})_{11} > 0$, so at most one sign change is possible. Setting $p_p = 1$ in (41) yields (31) as the necessary and sufficient condition for this not to occur in the feasible region $p_p \leq 1$. The special cases follow simply from (31).

E. Proof of Theorem 20

First consider the case $2 \leq j \leq W$. From (21) and (24)

$$(\mathbf{J}_{\text{SH}}^{-1})_{jj} \geq \frac{\theta_j}{p_p} \geq \frac{\theta_j}{p} \geq \frac{\theta_j}{p} - \frac{q}{p} \theta_j^2 = (\mathbf{J}_{\text{FS}}^{-1})_{jj}$$

since $p_f = p$ and $p_p = p_p(p) \leq p$ under both PPR and ESR.

Now consider $j = 1$. It is convenient to recall (23) and (21): $(\mathbf{J}_{\text{SH}}^{-1})_{11} = \theta_1 + \frac{1}{p_p} \sum_{k=2}^W q_p^{k-1} \theta_k$, and $(\mathbf{J}_{\text{FS}}^{-1})_{11} = \frac{1}{p_f} \theta_1 - \frac{q_f}{p_f} \theta_1^2$. It follows that $(\mathbf{J}_{\text{FS}}^{-1})_{11} \leq (\mathbf{J}_{\text{SH}}^{-1})_{11}$ when

$$p_p \leq \frac{p \sum_{k=2}^W q_p^{k-1} \theta_k}{\theta_1 (1 - \theta_1) (1 - p)}. \quad (42)$$

A sufficient condition implying (42) is obtained by using the lower bound $q_p \theta_2 \leq \sum_{k=2}^W q_p^{k-1} \theta_k$ and rearranging, yielding

$$p_p \leq \frac{p \theta_2}{\theta_1 (1 - \theta_1) (1 - p) + p \theta_2}. \quad (43)$$

Furthermore, since $p_p \leq p$, a more restrictive sufficient condition is given by replacing the l.h.s. with p , which reduces to $p(1-p)(\theta_1(1-\theta_1) - \theta_2) \leq 0$, which shows that for any $0 \leq p \leq 1$, if $\theta_2 \geq \theta_1(1-\theta_1)$, then (42) holds and $(\mathbf{J}_{\text{FS}}^{-1})_{11} \leq (\mathbf{J}_{\text{SH}}^{-1})_{11}$. The condition is satisfied if $W = 2$.

Now let θ be arbitrary and consider the small p (and hence small p_p) limit. Then (42) becomes $p_p \leq p/\theta_1$, which is always satisfied since $p_p \leq p$.

APPENDIX E MAXIMUM LIKELIHOOD ESTIMATORS

A. Proof of Theorem 22

The likelihood function for N_f flows is

$$f(\theta, N_f) = \prod_i f(j_i; \theta) = \prod_{j \geq 0} d_j^{M'_j}.$$

The MLE is the θ which maximizes the log-likelihood

$$l(\theta, N_f) = \sum_{j \geq 0} M'_j \log d_j$$

subject to the constraint $\sum_{k \geq 1} \theta_k = 1$, $\theta_k > 0$, $\forall k$. The optimization problem admits a feasible solution by the Bolzano-Weierstrass theorem [29, p. 517], since the log-likelihood function is concave and continuous, and optimization is performed over a compact, convex set. Furthermore, the solution obtained will be unique under our assumptions, since the Hessian of the log-likelihood is the Fisher information, which is positive definite given $0 < \theta_k < 1$ for all k .

Given the assumptions, the method of Lagrange multipliers would yield the optimal solution since strong duality holds

as the problem satisfies Slater's constraint qualification [30, Section 5.2.3, p. 226]. The Lagrangian is

$$\mathcal{L}(\boldsymbol{\theta}, \lambda, \boldsymbol{\nu}) = \sum_{j \geq 0} M'_j \log d_j - \lambda \left(\sum_{k \geq 1} \theta_k - 1 \right) - \boldsymbol{\nu}^T \boldsymbol{\theta},$$

where the vector $\boldsymbol{\nu}$ has elements $\nu_k \geq 0$ and $\lambda \in \mathbb{R}$. By differentiating with respect to θ_k , $\forall k$ and the multipliers,

$$\frac{\partial}{\partial \theta_k} \mathcal{L}(\boldsymbol{\theta}, \lambda, \boldsymbol{\nu}) = \sum_{j \geq k} \frac{M'_j}{d_j} b_{jk} - \lambda - \nu_k = 0,$$

$$\frac{\partial}{\partial \lambda} \mathcal{L}(\boldsymbol{\theta}, \lambda, \boldsymbol{\nu}) = 1 - \sum_{k \geq 1} \theta_k = 0,$$

$$\frac{\partial}{\partial \nu_k} \mathcal{L}(\boldsymbol{\theta}, \lambda, \boldsymbol{\nu}) = \theta_k = 0.$$

The second equation is just the equality constraint while the third yields a solution $\boldsymbol{\theta} = \mathbf{0}_W$, which lies on the boundary, yielding an unbounded solution (observed by substituting the solution into the likelihood function). That leaves the first equation, and by the Karush-Kuhn-Tucker condition, $\boldsymbol{\nu}^T \boldsymbol{\theta} = 0$ [30, Section 5.5.3, p. 243], implying that $\boldsymbol{\nu} = \mathbf{0}_W$ (our assumptions require that the original parameters $0 < \theta_k < 1$ for all k , hence the optimal must lie within the region where the constraints are inactive). Thus, we have, in matrix form,

$$\tilde{\mathbf{B}}^T \tilde{\mathbf{D}} \text{diag}(M'_1, \dots, M'_W) \mathbf{1}_W = \lambda \mathbf{1}_W - \frac{M'_0}{d_0} \mathbf{b}_0, \quad (44)$$

recalling that $\tilde{\mathbf{D}} = \text{diag}(d_1^{-1}, \dots, d_W^{-1})$.

We proceed to solve for λ using the equality constraint $\sum_{k \geq 1} \theta_k = 1$ and $\tilde{\mathbf{d}} = \tilde{\mathbf{B}} \boldsymbol{\theta}$, as follows, by multiplying both sides of (44) by $\boldsymbol{\theta}^T$ to obtain

$$\begin{aligned} \boldsymbol{\theta}^T \tilde{\mathbf{B}}^T \tilde{\mathbf{D}} \text{diag}(M'_1, \dots, M'_W) \mathbf{1}_W &= \lambda - \frac{M'_0}{d_0} \\ \tilde{\mathbf{d}}^T \tilde{\mathbf{D}} \text{diag}(M'_1, \dots, M'_W) \mathbf{1}_W &= \lambda - \frac{M'_0}{d_0} \\ \mathbf{1}_W^T \text{diag}(M'_1, \dots, M'_W) \mathbf{1}_W &= \lambda - \frac{M'_0}{d_0}, \end{aligned}$$

implying $\lambda = N_f$.

For methods that pick flows in an unbiased manner, $\mathbf{b}_0 = q \mathbf{1}_W$, and thus $\tilde{\mathbf{B}}^T \mathbf{1}_W = p \mathbf{1}_W$, implying $(\tilde{\mathbf{B}}^{-1})^T \mathbf{1}_W = p^{-1} \mathbf{1}_W$, therefore (44) reduces to

$$\tilde{\mathbf{B}}^T \tilde{\mathbf{D}} \text{diag}(M'_1, \dots, M'_W) \mathbf{1}_W = (N_f - M'_0) \mathbf{1}_W. \quad (45)$$

which can be rewritten as

$$\begin{aligned} \tilde{\mathbf{D}}^{-1} (\tilde{\mathbf{B}}^{-1})^T \mathbf{1}_W &= \frac{1}{N_f - M'_0} \text{diag}(M'_1, \dots, M'_W) \mathbf{1}_W \\ \tilde{\mathbf{B}} \boldsymbol{\theta} &= \frac{p}{\sum_{j=1}^W M'_j} [M'_1, \dots, M'_W]^T. \end{aligned}$$

The result follows.

Remark 32: In the conditional framework, expressing the log-likelihood function is difficult, due the fact that normalization of the likelihood involves division by random variables. However, the estimator above would, with high probability, be close to the actual MLE. The flow selection process is a Bernoulli process. The denominator, $\sum_{j=1}^W M'_j$ encapsulates information about N_f , because asymptotically, the deviation between $p N_f$ and $\sum_{j=1}^W M'_j$ is extremely small, a consequence of the concentration of Bernoulli samples around its mean.

This property is not found amongst other methods, such as PS, where samples are biased towards large flows.

B. Proof of Theorem 23

We begin with the optimization equation (44),

$$\tilde{\mathbf{B}}^T \tilde{\mathbf{D}} \text{diag}(M'_1, \dots, M'_W) \mathbf{1}_W + \frac{M'_0}{d_0} \mathbf{b}_0 = N_f \mathbf{1}_W.$$

Using properties of the sampling matrix for SH, we have

$$\tilde{\mathbf{D}} \text{diag}(M'_1, \dots, M'_W) \mathbf{1}_W + \frac{M'_0}{d_0} \cdot \frac{q}{p} \mathbf{e}_1 = N_f (\tilde{\mathbf{B}}^{-1})^T \mathbf{1}_W \quad (46)$$

$$\tilde{\mathbf{D}} \text{diag}(M'_0 + M'_1, \dots, M'_W) \mathbf{1}_W = N_f (\tilde{\mathbf{B}}^{-1})^T \mathbf{1}_W \quad (47)$$

$$\tilde{\mathbf{D}} \text{diag}(p(M'_0 + M'_1), \dots, M'_W) \mathbf{1}_W = N_f \mathbf{1}_W. \quad (48)$$

where in (46) we use the property $\mathbf{b}_0^T = \frac{q}{p} \tilde{\mathbf{B}}^T \mathbf{e}_1$, where \mathbf{e}_i is the canonical basis vector, in (47) we use $d_0 = \frac{q}{p} d_1$, and in (48) we use $(\tilde{\mathbf{B}}^{-1})^T \mathbf{1}_W = \text{diag}(p^{-1}, 1, \dots, 1) \mathbf{1}_W$. All these properties can be obtained by a straightforward evaluation using the sampling matrix. The final line reduces to

$$\tilde{\mathbf{B}} \boldsymbol{\theta} = \frac{1}{N_f} \cdot [p(M'_0 + M'_1), \dots, M'_W]^T,$$

proving the result.

The estimator is unbiased, as by taking the expectation, we obtain $\mathbb{E}[p(M'_0 + M'_1)] = N_f(p d_0 + p d_1) = N_f(p + q) d_1 = d_1$, by using the identity $d_1 = \frac{p}{q} d_0$ while clearly $\mathbb{E}[M'_j] = d_j$ for all $j \geq 2$. Thus $\mathbb{E}[\hat{\boldsymbol{\theta}}_{\text{SH}}] = \tilde{\mathbf{B}}_{\text{SH}}^{-1} \tilde{\mathbf{B}}_{\text{SH}} \boldsymbol{\theta} = \boldsymbol{\theta}$.

REFERENCES

- [1] N. Duffield, C. Lund, and M. Thorup, "Estimating Flow Distributions from Sampled Flow Statistics," *IEEE/ACM Trans. Networking*, vol. 13, no. 5, pp. 933–946, Oct 2005.
- [2] N. Hohn and D. Veitch, "Inverting Sampled Traffic," in *Proc. 2003 ACM SIGCOMM Internet Measurement Conference*, Miami, October 2003, pp. 222–233.
- [3] —, "Inverting Sampled Traffic," *IEEE/ACM Transactions on Networking*, vol. 14, no. 1, pp. 68–80, 2006.
- [4] B. Ribeiro, D. Towsley, T. Ye, and J. Bolot, "Fisher information on sampled packets: an application to flow size estimation," in *Proc. ACM/SIGCOMM Internet Measurement Conf.*, Rio de Janeiro, Oct 2006, pp. 15–26.
- [5] G. Varghese, *Network Algorithmics*. San Francisco: Elsevier/Morgan Kaufmann, 2005.
- [6] C. Estan and G. Varghese, "New directions in traffic measurement and accounting," *ACM Transactions on Computer Systems*, vol. 21, no. 3, pp. 270–313, August 2003.
- [7] P. Tune and D. Veitch, "Towards Optimal Sampling for Flow Size Estimation," in *Proc. ACM SIGCOMM Internet Measurement Conf.*, Vouliagmeni, Greece, Oct.20-22 2008, pp. 243–256.
- [8] L. Yang and G. Michailidis, "Estimation of flow lengths from sampled traffic," in *Proc. GLOBECOM*, San Francisco, CA, Nov 2006.
- [9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. John Wiley and Sons, Inc., 2006.
- [10] A. Hero, J. Fessler, and M. Usman, "Exploring estimator bias-variance tradeoffs using the Uniform CR bound," *IEEE Trans. Sig. Proc.*, vol. 44, no. 8, pp. 2026–2041, Aug 1996.
- [11] J. D. Gorman and A. O. Hero, "Lower bounds for parametric estimation with constraints," *IEEE Trans. Info. Th.*, vol. 26, no. 6, pp. 1285–1301, Nov 1990.
- [12] D. Harville, *Matrix Algebra from a Statistician's Perspective*. Springer-Verlag, 1997.
- [13] J. E. Strum, "Binomial matrices," *The Two Year College Mathematics Journal*, vol. 8, no. 5, pp. 260–266, November 1977.

- [14] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, 2nd ed., ser. Springer Texts in Statistics. Springer, 1998.
- [15] P. Tune and D. Veitch, "Fisher information in flow size distribution estimation: Technical report," Dept. E&EE, The University of Melbourne, Tech. Rep., 2008, copy available upon request.
- [16] D. Shah, S. Iyer, B. Prabhakar, and N. McKeown, "Maintaining statistics counters in line cards," *IEEE Micro*, vol. 22, no. 1, pp. 76–81, 2002.
- [17] S. Ramabhadran and G. Varghese, "Efficient implementation of a statistics counter architecture," *ACM SIGMETRICS Performance Evaluation Review*, vol. 31, no. 1, pp. 261–271, June 2003.
- [18] Q. Zhao, J. Xu, and Z. Liu, "Design of a novel statistics counter architecture with optimal space and time efficiency," *Proc. of ACM SIGMETRICS 2006*, vol. 34, no. 1, pp. 323–334, June 2006.
- [19] R. T. Rockafellar, *Convex Analysis*, ser. Princeton Landmarks in Mathematics and Physics. Princeton University Press, 1970.
- [20] C. Estan, K. Keyes, D. Moore, and G. Varghese, "Building a better netflow," in *Proc. ACM SIGCOMM 2004*, Portland, OR, Aug 2004.
- [21] NLANR, Leipzig-II Trace Data, <http://pma.nlanr.net/Special/leip2.html>.
- [22] —, Abilene-III Trace Data, <http://pma.nlanr.net/Special/ipls3.html>.
- [23] CoralReef, <http://www.caida.org/tools/measurement/coralreef/>.
- [24] S. Kay, *Fundamentals of Statistical Signal Processing, Volume I, Estimation Theory*. Prentice Hall, 1993.
- [25] M. Mitzenmacher and E. Upfal, *Probability and Computing*. Cambridge University Press, 2005.
- [26] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*, 2nd ed. Wiley Interscience, 2008.
- [27] F. Zhang, *Matrix Theory: Basic Results and Techniques*. Springer-Verlag, 1999.
- [28] R. Zamir, "A proof of the Fisher information inequality via a data processing argument," *IEEE Transactions on Information Theory*, vol. 44, no. 3, pp. 1246–1250, May 1998.
- [29] A. E. Taylor and W. R. Mann, *Advanced Calculus*, 3rd ed. John Wiley and Sons, 1983.
- [30] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.



Paul Tune (M'09) received the B.E. (Hon.) and B.Sc. degrees in Electrical and Electronics Engineering and Computer Science respectively, both in 2005, from the University of Melbourne, Australia. He has recently completed a Ph.D. degree at the ARC Special Center for Ultra-Broadband Information Networks (CUBIN) at the University of Melbourne, Australia. His research interests are in communication networks, particularly in traffic inversion and sampling, information theory, statistics and signal processing, specifically

sparse signal recovery.



Darryl Veitch (F'10) completed a BSc. Hons. at Monash University, Australia (1985) and a mathematics Ph.D. from Cambridge, England (1990). He worked at TRL (Telstra, Melbourne), CNET (France Telecom, Paris), KTH (Stockholm), INRIA (Sophia Antipolis, France), Bellcore (New Jersey), RMIT (Melbourne) and EMulab and CUBIN at The University of Melbourne, where he is a Professorial Research Fellow. His research interests include traffic modelling, parameter estimation, active measurement, traffic sampling,

and clock synchronization.

